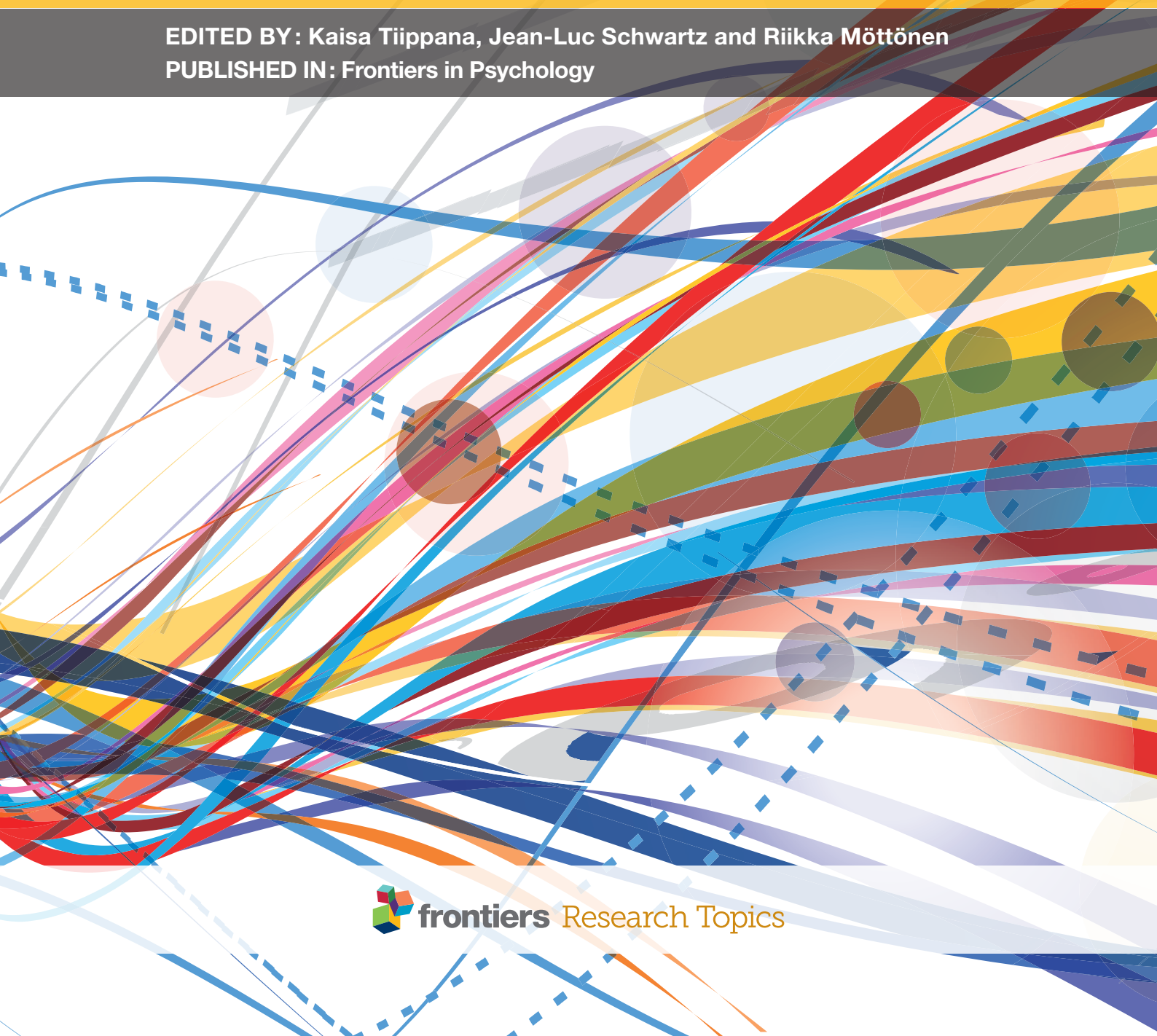
MULTISENSORY AND SENSORIMOTOR INTERACTIONS IN SPEECH PERCEPTION

EDITED BY: Kaisa Tiippana, Jean-Luc Schwartz and Riikka Möttönen
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-548-0
DOI 10.3389/978-2-88919-548-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

MULTISENSORY AND SENSORIMOTOR INTERACTIONS IN SPEECH PERCEPTION

Topic Editors:
Kaisa Tiippana, University of Helsinki, Finland
Jean-Luc Schwartz, Grenoble University, France
Riikka Möttönen, University of Oxford, United Kingdom

Speech is multisensory since it is perceived through several senses. Audition is the most important one as speech is mostly heard. The role of vision has long been acknowledged since many articulatory gestures can be seen on the talker's face. Sometimes speech can even be felt by touching the face. The best-known multisensory illusion is the McGurk effect, where incongruent visual articulation changes the auditory percept. The interest in the McGurk effect arises from a major general question in multisensory research: How is information from different senses combined? Despite decades of research, a conclusive explanation for the illusion remains elusive. This is a good demonstration of the challenges in the study of multisensory integration.

Speech is special in many ways. It is the main means of human communication, and a manifestation of a unique language system. It is a signal with which all humans have a lot of experience. We are exposed to it from birth, and learn it through development in face-to-face contact with others. It is a signal that we can both perceive and produce. The role of the motor system in speech perception has been debated for a long time. Despite very active current research, it is still unclear to which extent, and in which role, the motor system is involved in speech perception. Recent evidence shows that brain areas involved in speech production are activated during listening to speech and watching a talker's articulatory gestures. Speaking involves coordination of articulatory movements and monitoring their auditory and somatosensory consequences. How do auditory, visual, somatosensory, and motor brain areas interact during speech perception? How do these sensorimotor interactions contribute to speech perception?

It is surprising that despite a vast amount of research, the secrets of speech perception have not yet been solved. The multisensory and sensorimotor approaches provide new opportunities in solving them. Contributions to the research topic are encouraged for a wide spectrum of research on speech perception in multisensory and sensorimotor contexts, including novel experimental findings ranging from psychophysics to brain imaging, theories and models, reviews and opinions.

Citation: Tiippana, K., Schwartz, J. L., Möttönen, R., eds. (2015). Multisensory and Sensorimotor Interactions in Speech Perception. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-548-0

Table of Contents

frontiers in Psychology

EDITORIAL
published: 20 April 2015
doi: 10.3389/fpsyg.2015.00458

Multisensory and sensorimotor interactions in speech perception

Kaisa Tiippana[1]*, *Riikka Möttönen*[2] and *Jean-Luc Schwartz*[3]

[1] *Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland,* [2] *Department of Experimental Psychology, University of Oxford, Oxford, UK,* [3] *Grenoble Images Parole Signal Automatique-Lab, Speech and Cognition Department, Centre National de la Recherche Scientifique, Grenoble University, Grenoble, France*

Keywords: audiovisual, cognitive disorders, learning, McGurk effect, multisensory, sensorimotor, somatosensory, speech perception

This research topic presents speech as a natural, well-learned, multisensory communication signal, processed by multiple mechanisms. Reflecting the general status of the field, most articles focus on audiovisual speech perception and many utilize the McGurk effect, which arises when discrepant visual and auditory speech stimuli are presented (McGurk and MacDonald, 1976). Tiippana (2014) argues that the McGurk effect can be used as a proxy for multisensory integration provided it is not interpreted too narrowly.

Several articles shed new light on audiovisual speech perception in special populations. It is known that individuals with autism spectrum disorder (ASD, e.g., Saalasti et al., 2012) or language impairment (e.g., Meronen et al., 2013) are generally less influenced by the talking face than peers with typical development. Here Stevenson et al. (2014) propose that a deficit in multisensory integration could be a marker of ASD, and a component of the associated deficit in communication. However, three studies suggest that integration is not deficient in some communication disorders. Irwin and Brancazio (2014) show that children with ASD looked less at the mouth region, resulting in poorer visual speech perception and consequently weaker visual influence. Leybaert et al. (2014) report that children with specific language impairment recognized visual and auditory speech less accurately than their controls, affecting audiovisual speech perception, while audiovisual integration *per se* seemed unimpaired. In a similar vein, adult patients with aphasia showed unisensory deficits but still integrated audiovisual speech information (Andersen and Starrfelt, 2015).

Multisensory information can influence response accuracy and processing speed (e.g., Molholm et al., 2002; Klucharev et al., 2003). Scarbel et al. (2014) show that oral responses to speech in noise were faster but less accurate than manual responses, suggesting that oral responses are planned at an earlier stage than manual responses. Sekiyama et al. (2014) show that older adults were more influenced by visual speech than younger adults and correlated this fact to their slower reaction times to auditory stimuli. Altieri and Hudock (2014) report variation in reaction time and accuracy benefits for audiovisual speech in hearing-impaired observers, emphasizing the importance of individual differences in integration. Finally, Heald and Nusbaum (2014) show that when there were two possible talkers instead of just one, audiovisual information appeared to distract the observer from the task of word recognition and slowed down their performance. This finding demonstrates that multisensory stimulation does not always facilitate performance.

While multisensory stimulation is thought to be beneficial for learning (Shams and Seitz, 2008), evidence for this is still scarce. In the current research topic, the overall utility of multisensory learning is brought under question. In a paradigm training to associate novel words and pictures, Bernstein et al. (2014) show no benefit of audiovisual presentation compared with auditory presentation for normal hearing individuals, and even a degradation for adults with hearing impairment. In a study of cued speech, i.e., specific hand-signs for different speech sounds, Bayard et al. (2014) demonstrate that individuals with hearing impairment used the visual cues differently from their controls, even though both groups were experts in cued speech. Kelly et al. (2014)

OPEN ACCESS

Edited and reviewed by:
Manuel Carreiras, Basque Center on Cognition, Brain and Language, Spain

***Correspondence:**
Kaisa Tiippana, kaisa.tiippana@helsinki.fi

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *27 March 2015*
Accepted: *30 March 2015*
Published: *20 April 2015*

Citation:
Tiippana K, Möttönen R and Schwartz J-L (2015) Multisensory and sensorimotor interactions in speech perception. Front. Psychol. 6:458. doi: 10.3389/fpsyg.2015.00458

show that when normal hearing adults learned words in a foreign language, viewing or producing hand gestures accompanying audiovisual speech did not affect the outcome. Lee and Noppeney (2014) show that musicians had a narrower audiovisual temporal integration window for music, and to a smaller extent also for speech, implying that the effect transfers from the practiced music stimuli also to other stimulus types. Together, these findings suggest that long-term training and active use may be requisites for multisensory information to be useful in learning speech.

Neurophysiological correlates of audiovisual speech perception were addressed in the research topic. By using electroencephalography (EEG) it was shown that attention (Alsius et al., 2014) and stimulus context (Ganesh et al., 2014) affected early event-related potentials (ERPs) to audiovisual speech. This provides further evidence that audiovisual interactions are not completely automatic. By using functional magnetic resonance imaging, Erickson et al. (2014) demonstrate a subdivision of posterior superior temporal areas for integrating congruent vs. incongruent audiovisual speech, and Callan et al. (2014) show that different regions in the premotor cortex were involved in unisensory-to-articulatory mapping and audiovisual integration.

Interactions between auditory and motor brain areas during auditory speech perception were also investigated. By using magnetoencephalography, Alho et al. (2014) demonstrate that connectivity between auditory and motor areas increased from passive listening to clear speech to listening to speech in noise, and that the strength of this connectivity was positively correlated with the accuracy of syllable identification. Moreover, analyses of EEG oscillations revealed that alpha and beta rhythms generated in the sensorimotor and auditory areas were modulated during syllable discrimination tasks (Bowers et al., 2014; Jenson et al., 2014). By using theta-burst transcranial magnetic stimulation, Rogers et al. (2014) show that disrupting the lip area of the motor cortex impaired discrimination of lip-articulated speech sounds from sounds not articulated on the lips. The involvement of the motor processes is often considered to make speech perception "special," i.e., essentially different from perception of non-speech stimuli. However, this remains a highly controversial view. Carbonell and Lotto (2014) claim that speech should not be considered special amongst other stimuli with regards to multisensory integration.

Somatosensory information can also influence speech perception. Ito et al. (2014) used EEG to study how stretching the skin on both sides of the mouth influences processing of speech sounds, and displayed auditory-somatosensory interaction that was sensitive to intersensory timing. In another EEG study, Treille et al. (2014) report that haptic exploration of the talker's face during speech perception modulated ERPs. These findings confirm that auditory-somatosensory interactions contribute to speech processing.

The current research topic shows that speech can be perceived via multiple senses and that speech perception relies on sophisticated unisensory, multisensory and sensorimotor mechanisms. Multisensory information can facilitate perception and learning of speech. Still, there is great variation in multisensory perception and integration in both typical and special populations at different ages, which should be studied further in the future.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) (Grant Agreement no. 339152, Speech Unit(e)s. Principal Investigator JS) Medical Research Council U.K. (Career Development Fellowship to RM) and the University of Helsinki (research grant to KT).

References

Alho, J., Lin, F. H., Sato, M., Tiitinen, H., Sams, M., and Jääskeläinen, I. P. (2014). Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. *Front. Psychol.* 5:394. doi: 10.3389/fpsyg.2014.00394

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., and Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Front. Psychol.* 5:727. doi: 10.3389/fpsyg.2014.00727

Altieri, N., and Hudock, D. (2014). Hearing impairment and audiovisual speech integration ability: a case study report. *Front. Psychol.* 5:678. doi: 10.3389/fpsyg.2014.00678

Andersen, T. S., and Starrfelt, R. (2015). Audiovisual integration of speech in a patient with Broca's Aphasia. *Front. Psychol.* 6:435. doi: 10.3389/fpsyg.2015.00435

Bayard, C., Colin, C., and Leybaert, J. (2014). How is the McGurk effect modulated by Cued Speech in deaf and hearing adults? *Front. Psychol.* 5:416. doi: 10.3389/fpsyg.2014.00416

Bernstein, L. E., Eberhardt, S. P., and Auer, E. T. (2014). Audiovisual spoken word training can promote or impede auditory-only perceptual learning: results from prelingually deafened adults with late-acquired cochlear implants and normal-hearing adults. *Front. Psychol.* 5:934. doi: 10.3389/fpsyg.2014.00934

Bowers, A. L., Saltuklaroglu, T., Harkrider, A., Wilson, M., and Toner, M. A. (2014). Dynamic modulation of shared sensory and motor cortical rhythms mediates speech and non-speech discrimination performance. *Front. Psychol.* 5:366. doi: 10.3389/fpsyg.2014.00366

Callan, D. E., Jones, J. A., and Callan, A. (2014). Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. *Front. Psychol.* 5:389. doi: 10.3389/fpsyg.2014.00389

Carbonell, K. M., and Lotto, A. J. (2014). Speech is not special... again. *Front. Psychol.* 5:427. doi: 10.3389/fpsyg.2014.00427

Erickson, L. C., Zielinski, B. A., Zielinski, J. E., Liu, G., Turkeltaub, P. E., Leaver, A. M., et al. (2014). Distinct cortical locations for integration of audiovisual speech and the McGurk effect. *Front. Psychol.* 5:534. doi: 10.3389/fpsyg.2014.00534

Ganesh, A. C., Berthommier, F., Vilain, C., Sato, M., and Schwartz, J.-L. (2014). A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Front. Psychol.* 5:1340. doi: 10.3389/fpsyg.2014.01340

Heald, S., and Nusbaum, H. C. (2014). Talker variability in audiovisual speech perception. *Front. Psychol.* 5:698. doi: 10.3389/fpsyg.2014.00698

Irwin, J., and Brancazio, L. (2014). Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. *Front. Psychol.* 5:397. doi: 10.3389/fpsyg.2014.00397

Ito, T., Gracco, V. L., and Ostry, D. J. (2014). Temporal factors affecting somatosensory-auditory interactions in speech processing. *Front. Psychol.* 5:1198. doi: 10.3389/fpsyg.2014.01198

Jenson, D., Bowers, A. L., Harkrider, A., Thornton, D., Cuellar, M., and Saltuklaroglu, T. (2014). Temporal dynamics of sensorimotor integration in speech perception and production: independent component analysis of EEG data. *Front. Psychol.* 5:656. doi: 10.3389/fpsyg.2014.00656

Kelly, S., Hirata, Y., Manansala, M., and Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Front. Psychol.* 5:673. doi: 10.3389/fpsyg.2014.00673

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res. Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Lee, H. L., and Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech and music. *Front. Psychol.* 5:868. doi: 10.3389/fpsyg.2014.00868

Leybaert, J., Macchi, L., Huyse, A., Champoux, F., Bayard, C., Colin, C., et al. (2014). Atypical audio-visual speech perception and McGurk effects in children with specific language impairment. *Front. Psychol.* 5:422. doi: 10.3389/fpsyg.2014.00422

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature,* 264, 746–748. doi: 10.1038/264746a0

Meronen, A., Tiippana, K., Westerholm, J., and Ahonen, T. (2013). Audiovisual speech perception in children with developmental language disorder in degraded listening conditions. *J. Speech Lang. Hear. Res.* 56, 211–221. doi: 10.1044/1092-4388(2012/11-0270)

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research,* 14, 115–128. doi: 10.1016/S0926-6410(02)00066-6

Rogers, J. C., Möttönen, R., Boyles, R., and Watkins, K. E. (2014). Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex. *Front. Psychol.* 5:754. doi: 10.3389/fpsyg.2014.00754

Saalasti, S., Kätsyri, J., Tiippana, K., Laine-Hernandez, M., von Wendt, L., and Sams, M. (2012). Audiovisual speech perception and eye gaze behavior of adults with Asperger Syndrome. *J. Autism Dev. Disord.* 42, 1606–1615. doi: 10.1007/s10803-011-1400-0

Scarbel, L., Beautemps, D., Schwartz, J.-L., and Sato, M. (2014). The shadow of a doubt ? Evidence for perceptuo-motor linkage during auditory and audiovisual close shadowing. *Front. Psychol.* 5:568. doi: 10.3389/fpsyg.2014.00568

Sekiyama, K., Soshi, T., and Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Front. Psychol.* 5:323. doi: 10.3389/fpsyg.2014.00323

Shams, L., and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends Cogn. Sci.* 12, 411–417. doi: 10.1016/j.tics.2008.07.006

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., and Wallace, M. T. (2014). The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. *Front. Psychol.* 5:379. doi: 10.3389/fpsyg.2014.00379

Tiippana, K. (2014). What is the McGurk effect? *Front. Psychol.* 5:725. doi: 10.3389/fpsyg.2014.00725

Treille, A., Vilain, C., and Sato, M. (2014). The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. *Front. Psychol.* 5:420. doi: 10.3389/fpsyg.2014.00420

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 06 May 2014
doi: 10.3389/fpsyg.2014.00394

Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization

Jussi Alho[1]*, Fa-Hsuan Lin[1,2], Marc Sato[3], Hannu Tiitinen[1], Mikko Sams[1] and Iiro P. Jääskeläinen[1,4,5]*

[1] Brain and Mind Laboratory, Department of Biomedical Engineering and Computational Science (BECS), School of Science, Aalto University, Espoo, Finland
[2] Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan
[3] Gipsa-Lab, Department of Speech and Cognition, French National Center for Scientific Research and Grenoble University, Grenoble, France
[4] MEG Core, Aalto NeuroImaging, School of Science, Aalto University, Espoo, Finland
[5] AMI Centre, Aalto NeuroImaging, School of Science, Aalto University, Espoo, Finland

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Alessandro D'Ausilio, Italian Institute of Technology, Italy
Ediz Sohoglu, University College London, UK

***Correspondence:**
Jussi Alho and Iiro P. Jääskeläinen, Brain and Mind Laboratory, Department of Biomedical Engineering and Computational Science (BECS), School of Science, Aalto University, P.O. Box 12200, FI-00076 Aalto, Finland
e-mail: jussi.alho@aalto.fi; iiro.jaaskelainen@aalto.fi

The cortical dorsal auditory stream has been proposed to mediate mapping between auditory and articulatory-motor representations in speech processing. Whether this sensorimotor integration contributes to speech perception remains an open question. Here, magnetoencephalography was used to examine connectivity between auditory and motor areas while subjects were performing a sensorimotor task involving speech sound identification and overt repetition. Functional connectivity was estimated with inter-areal phase synchrony of electromagnetic oscillations. Structural equation modeling was applied to determine the direction of information flow. Compared to passive listening, engagement in the sensorimotor task enhanced connectivity within 200 ms after sound onset bilaterally between the temporoparietal junction (TPJ) and ventral premotor cortex (vPMC), with the left-hemisphere connection showing directionality from vPMC to TPJ. Passive listening to noisy speech elicited stronger connectivity than clear speech between left auditory cortex (AC) and vPMC at ∼100 ms, and between left TPJ and dorsal premotor cortex (dPMC) at ∼200 ms. Information flow was estimated from AC to vPMC and from dPMC to TPJ. Connectivity strength among the left AC, vPMC, and TPJ correlated positively with the identification of speech sounds within 150 ms after sound onset, with information flowing from AC to TPJ, from AC to vPMC, and from vPMC to TPJ. Taken together, these findings suggest that sensorimotor integration mediates the categorization of incoming speech sounds through reciprocal auditory-to-motor and motor-to-auditory projections.

Keywords: magnetoencephalography, MEG, speech perception, dorsal stream, sensorimotor integration, premotor cortex

INTRODUCTION

Current theories propose that speech is cortically processed by the ventral and dorsal auditory streams (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). While the ventral stream processes acoustic-phonetic features of speech, the dorsal stream has been suggested to mediate mapping between auditory and articulatory-motor representations (Hickok et al., 2011; Rauschecker, 2011). Whether this sensorimotor integration contributes to the perception of others' speech remains debated (Cappa and Pulvermuller, 2012; Hickok, 2012; Schwartz et al., 2012).

As the speech signal has high variability and complex composition of acoustic features, it has been suggested that the listener's internal articulatory knowledge might be important in the categorization of incoming speech sounds (Liberman et al., 1967; Liberman and Mattingly, 1985; Davis and Johnsrude, 2007; Schwartz et al., 2012). Experimental support for such motor contribution is provided by findings showing that disturbing the left premotor cortex (PMC) or lip/tongue areas in the primary motor cortex (MC) with transcranial magnetic stimulation (TMS) results in impaired speech sound identification and discrimination (Meister et al., 2007; Möttönen and Watkins, 2009; Sato et al., 2009; D'Ausilio et al., 2011; Grabski et al., 2013). Möttönen et al. (2013) further demonstrated that the TMS-induced disruption of articulatory-motor cortex impairs also automatic speech sound discrimination (i.e., in the absence of behavioral tasks and without explicit attention directed to the speech sounds). In a related study, Chevillet et al. (2013) observed, using a functional magnetic resonance imaging (fMRI) adaptation paradigm, automatic phoneme category selectivity in the left PMC that correlated positively with behavioral categorization performance.

Further supporting the sensorimotor nature of speech perception, a study applying concurrent magnetoencephalography (MEG) and electroencephalography (EEG) with Granger causation analyzes found that activation in the posterior superior temporal gyrus (pSTG) was influenced by activation in dorsal PMC (dPMC) during perception of coarticulated speech, thus suggesting that articulatory processes directly mediate speech perception (Gow and Segawa, 2009). An fMRI study demonstrated that speech motor areas, in particular the ventral PMC (vPMC),

were more strongly activated by non-native compared to native phonemes, which can be interpreted as being caused by the motor system repeatedly iterating in order to find the best match for the unfamiliar acoustic input among candidate phonemic categorizations (Wilson and Iacoboni, 2006). A similar process can be expected in case of degraded native speech, as it has been shown that degraded compared to clear speech elicits enhanced responses in motor areas, including the inferior frontal gyrus (IFG) and PMC (e.g., Davis and Johnsrude, 2003). Relatedly, simultaneous MEG and EEG recordings demonstrated that perceptual clarity of degraded speech was enhanced by prior knowledge of speech content and associated with activity in the IFG that preceded activity changes in the STG, therefore suggesting that prior knowledge is integrated with speech inputs through top-down predictions from the speech motor areas to lower-level sensory cortex (Sohoglu et al., 2012).

Compatible with these studies, our recent MEG study with minimum-norm estimate (MNE) -based source modeling showed that activity in the left PMC was amplified at ∼200 ms after sound onset when subjects were to identify and repeat the presented speech sound compared to passive listening, with the effect being stronger when the sounds were masked by acoustic noise compared to clear speech (Alho et al., 2012). Also, the left PMC activity at ∼100 ms after sound onset correlated positively with speech sound identification accuracy. However, these findings alone do not answer the question whether performance in such sensorimotor task involves reciprocal auditory-to-motor and motor-to-auditory projections, which have been hypothesized to be crucial in constraining the interpretation of incoming acoustic speech information with complementary articulatory information (Schwartz et al., 2012). According to a recent dual-pathway model of auditory cortical processing, speech sounds are processed hierarchically in the ventral stream from the auditory cortex (AC) to the category-invariant inferior frontal cortex (IFC), transformed into articulatory representations in the vPMC, and finally transmitted to the temporoparietal junction (TPJ) as an efference copy (Rauschecker and Scott, 2009; Rauschecker, 2011). In this model, processing in the dorsal stream proceeds from the AC to the TPJ, where a quick sketch of sensory event information is compared with the efference copy of the activated articulatory-motor plans. Tentatively, such sensorimotor integration could be enabled by oscillatory synchrony, i.e., rhythmic millisecond-range temporal correlations of neuronal activity (Womelsdorf et al., 2007; Singer, 2009). Previous MEG and EEG studies have revealed that the level of inter-areal phase synchrony within the alpha (8–14 Hz), beta (14–30 Hz) and gamma (30–80 Hz) frequency bands correlates with various perceptual, attention, and working memory task performances (Kujala et al., 2007; Palva et al., 2010; Hipp et al., 2011; Kveraga et al., 2011; Huang et al., 2014), therefore supporting the hypothesis that coordinated operation between task-relevant brain regions is reflected as strengthened oscillatory synchrony (for a review, see Palva and Palva, 2012).

Here, we analyzed our previously published MEG dataset (Alho et al., 2012) to estimate functional connectivity among speech-relevant brain areas while subjects were performing a sensorimotor integration task involving speech sound identification and overt repetition. We utilized the increased spatiotemporal accuracy provided by MRI-based MNEs (Lin et al., 2006) to estimate inter-areal neural synchrony. Continuous wavelet transform of single-trial data was applied to reveal the phase dynamics of ongoing neural activity as a function of time and frequency. The level of phase synchrony was quantified with weighted phase lag index (WPLI; Vinck et al., 2011). In addition, directionality of information flow was estimated with structural equation modeling (SEM; Penny et al., 2004). We hypothesized that the neural synchrony between auditory and motor areas within 200 ms after sound onset is (1) enhanced when one is engaged in the sensorimotor task compared to passive listening; (2) enhanced when the sounds are masked by acoustic noise compared to clear speech; and (3) positively correlated with the speech sound identification accuracy.

MATERIALS AND METHODS

SUBJECTS

Twenty-two healthy individuals with self-reported normal hearing participated in the study. Two subjects were excluded from the analyses due to low signal-to-noise ratio (SNR), resulting in a final sample size of 20 subjects (18 right-handed, age range 21–58 years, mean ± SD age: 27.4 ± 8.0 years). All except one (Italian) were native speakers of Finnish. Informed consent was obtained from all subjects. The experiment was approved by the Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa.

STIMULI AND TASK

The stimuli were /pa/ and /ta/ syllable sounds articulated by a male native Finnish speaker and presented either as intact or embedded in noise. Five individual clearly articulated /pa/ and /ta/ tokens were selected, scaled to 68 dB, and cut at 100 ms preceding and following the detected consonantal burst. Thus, the duration of the spoken syllable was 100 ms. Noisy speech stimuli were created by masking the syllables with Gaussian pink noise. The masks had a 5-ms rise-decay envelope, were de-emphasized to better match the frequency spectrum of /pa/ and /ta/ syllables (at −6 dB/oct), and were simultaneously presented from the beginning to the end of the syllable with SNR of + 5 dB. A forced-choice identification test with a subset of six subjects was conducted to ensure appropriate syllable identification accuracy at this SNR level (i.e., 77% correct responses).

The stimuli were presented in four different conditions: passive perception; perception followed by overt repetition; perception followed by covert repetition; and perception followed by overt imitation. In the active conditions, the subjects' task was to identify the syllable as either /pa/ or /ta/, wait for a visual cue, and reproduce it accordingly. The overt imitation task differed from the overt repetition in that the reproduction of the target syllable was to be done by imitating the pitch of the stimulus sound. The covert repetition was to take place covertly without any articulatory movements or sound production.

Each condition comprised 300 trials (75 intact /pa/ + 75 intact /ta/ + 75 noisy /pa/ + 75 noisy /ta/) presented with (1) a randomly varying 1–1.5 s prestimulus baseline for perception, (2) randomized auditory stimulus presentation (/pa/ or /ta/), (3) a baseline for repetition of the syllable (300–800 ms after stimulus

offset), and (4) a visual cue to repeat (black fixation cross turning briefly to red; 2–2.2 s). Thus, the total duration of the trial was 6 s, with interstimulus interval (ISI) varying between 5.5 and 6.5 s, and the interval between the onset of the auditory stimulus and the subsequent visual cue to repeat varying between 0.5 and 1 s (**Figure 1**). The measurement time per condition totaled to ∼30 min, which was divided into two ∼15 min blocks to prevent fatigue. The measurements were divided on 2 days, with the passive listening and overt repetition conditions on the first day, and covert repetition and imitation conditions on the second day. The order of the conditions was kept fixed to reduce the possibility of the performance in the less demanding tasks being affected by the experience from the more demanding tasks (e.g., to reduce the subjects' disposition to covertly rehearse the presented stimuli in the passive listening condition or to imitate when natural repetition was required). The covert repetition and imitation conditions were not included in the analyses of the present study. The auditory stimuli were presented via a panel loudspeaker with an approximate 65-dB sound level. All stimuli were delivered with Presentation software (v10.1, Neurobehavioral systems).

DATA RECORDING

The MEG data were acquired with a whole-head 306-channel neuromagnetometer (VectorView, Elekta-Neuromag, Finland) of the MEG Core of Aalto NeuroImaging infrastructure at Aalto University. The device was situated in a magnetically shielded room, with a three-layer μ-metal and aluminum cover to attenuate effects of outside magnetic fields, and an additional active noise-cancelation system.

Before each MEG recording session, locations of four head position indicator (HPI) coils attached to the scalp were recorded with respect to three anatomical landmark points (nasion and two preauricular points) using a 3-D digitizer (Isotrak, Polhemus, Colchester, VT, USA). Additional scalp surface points (≈30) were digitized to facilitate coregistration with anatomical magnetic resonance (MR) images. To detect eye blinks and movements, an electro-oculogram (EOG) channel was recorded with electrodes placed below and on the outer canthus of the left eye. The MEG signals were band-pass filtered at 0.03–200 Hz and digitized at a sampling frequency of 2000 Hz. The individual MR images were acquired with a 3T GE Signa scanner (GE Healthcare Ltd., Chalfont St Giles, UK) of the AMI Center of Aalto NeuroImaging infrastructure at Aalto University.

For subsequent identification of the subjects' repetitions, microphone recordings with 22.05 kHz sampling rate together with electromyographic (EMG) channels with electrodes placed on three specific articulators (sternohyoid, orbicularis oris superior, and masseter) were recorded. The EMG responses were used also to control for the presence of any covert articulations that might have occurred after the perception of the syllables (i.e., before the onset of the cued reproduction task).

MEG SOURCE ESTIMATION

The MEG data were processed and analyzed with the MNE software package (Gramfort et al., 2014). The data were first downsampled to 1000 Hz and screened for artifacts. Epochs from 200 ms preceding and 500 ms following the stimulus onset were processed separately for the stimulus types. Non-functioning (i.e., flat) channels and trials with the epochs exceeding 3000 fT/cm amplitude (measured with respect to a 200-ms prestimulus baseline) in the MEG channels or 150 μV in the EOG channel were rejected from further analyses, resulting in an average of ∼120 trials/condition/stimulus type.

Source modeling was performed by computing MNEs (Hämäläinen and Ilmoniemi, 1994) from MRI-constrained MEG data. For this purpose, a single-compartment boundary element model (BEM; Hämäläinen and Sarvas, 1989) was constructed from the structural MRI and used as a forward model to constrain MEG source locations to the cortex. The source current strengths at each source location for each time point were estimated with the anatomically constrained linear estimation approach (Dale et al., 2000). To this end, an inverse operator was calculated with the help of a noise covariance matrix estimated from the filtered single-trial 200-ms prestimulus baselines. For visualizing the mean evoked activity on the cortical surface, dynamic statistical parametric map (dSPM) estimates were generated (Dale et al., 2000). As a measure of signal-to-noise (derived through normalizing the MNE by the noise sensitivity at each cortical location), dSPM indicates the locations with MNE amplitudes above the noise level. Since individual MRI-images were not available for six subjects, a FreeSurfer average brain was applied as a surrogate in these subjects (by aligning the individual fiducial points to the fiducial points of the average head).

REGIONS-OF-INTEREST (ROIs)

The inter-areal phase synchrony of the source data was investigated between ROIs. Considering that the MNE source estimation provides an underdetermined solution to the inverse problem (i.e., 306 measurement sensors to ∼7000 unknown source dipoles), five large anatomical regions per hemisphere were first selected on the basis of our a priori hypothesis by merging the labels of relevant gyri and sulci that resulted from the automatic anatomical

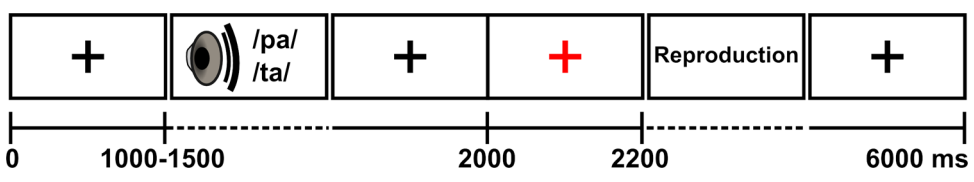


FIGURE 1 | Experimental procedure. Adapted from Alho et al. (2012).

parcellation (Destrieux et al., 2010): AC (comprising the superior temporal gyrus and sulcus), TPJ (comprising supramarginal gyrus, angular gyrus, and planum temporale), pIFG/vPMC (comprising the pars opercularis of the IFG and the inferior part of the precentral sulcus), dPMC (comprising the superior part of the precentral sulcus), and MC (comprising the central sulcus). Functional constraints were then applied to these anatomical regions by selecting only the subregions where the group-average dSPM activations exceeded a threshold value of 4 (F-statistic) at any time between 50 and 200 ms (see Statistical analysis for the selection criteria of the analysis time window). For minimizing bias (Kriegeskorte et al., 2009), the stimulus types and conditions used for the functional constraints between different analyses were as follows: noisy stimuli in the passive listening condition for the correlation tests between neural synchrony and syllable identification accuracy; combined noisy and intact stimuli in the passive listening condition for analyzing changes in neural synchrony between noisy and clear speech; and combined noisy and intact stimuli in combined passive and active listening (i.e., overt repetition) conditions for analyzing changes in neural synchrony between passive and active listening. The ROIs were defined on the FreeSurfer average brain (**Figure 2**) and morphed onto the individual surfaces with an automatic spherical morphing procedure (Fischl et al., 1999).

PHASE SYNCHRONY ESTIMATION

Single-trial raw (0.03–200 Hz) MNE currents from −200 to +500 ms were baseline corrected (with respect to the 200 ms prestimulus period), averaged over the source locations to obtain a time course for each ROI (by only keeping the radial components and applying sign-flips to reduce signal cancellations), and submitted to the phase synchrony analysis. Trials counts between conditions were equalized for reducing bias.

Phase synchrony between ROIs was estimated by computing a WPLI (Vinck et al., 2011) across trials for every time and frequency point. WPLI was chosen as a measure for its low sensitivity to the volume conductor effect (i.e., artificial synchrony caused by mixing of neuronal signals). This attribute is based on the idea that non-zero phase lag between two time courses is not caused by volume conduction from a common source, but rather by actual communication between brain structures through a physical medium, which is bound to have a delay (or a non-zero phase lag). The WPLIs were obtained by first filtering the ROI time courses with a continuous Morlet wavelet transform into 25 center frequencies from 8–80 Hz with 3 Hz steps (wavelet width varying from 1.1 at lowest frequency to 11.4 cycles at highest frequency). The non-zero phase lag interdependencies were then estimated, for a particular frequency, by weighting the contribution of observed phase leads and lags by the magnitude of the imaginary component of the cross-spectrum between each pair of ROIs (Vinck et al., 2011). WPLI-values range from 0 to 1, with 0 indicating random distribution of phase and 1 indicating constant (non-zero lag) phase difference across trials.

Statistical analysis

Spearman rank correlation test was applied to examine correlations between neural synchrony and syllable identification accuracy. For assessing changes in neural synchrony between active and passive listening, and their interaction with noisy vs. clear speech, a two-way repeated measures analysis of variance (ANOVA) was conducted. Changes in neural synchrony between noisy and clear speech was analyzed with one-way ANOVA in the passive condition to avoid the possible confounding effect caused by subjects covertly rehearsing the presented syllable while waiting for the visual cue in the active listening condition. As it has been shown that acoustic-phonetic features of speech modulate auditory cortical activity from 50 ms onwards and that the access to phonological categories occurs at ∼150 ms after stimulus onset (for a review, see Salmelin, 2007), a time range of 50–200 ms was selected for the analyses. Restricting the analysis to early latencies also decreases the likelihood that the phase synchrony effects might be due to speech preparation after subjects have identified the auditory target. Within the analysis range, the WPLIs were averaged into 10-ms time windows. The p-values were FDR-corrected for multiple ROI connection × time × frequency point comparisons (Benjamini et al., 2001).

To control for the possibility that the phase synchrony effects could be explained by the regions independently synchronizing to the stimulus onset (i.e., phase resetting by stimulus-evoked

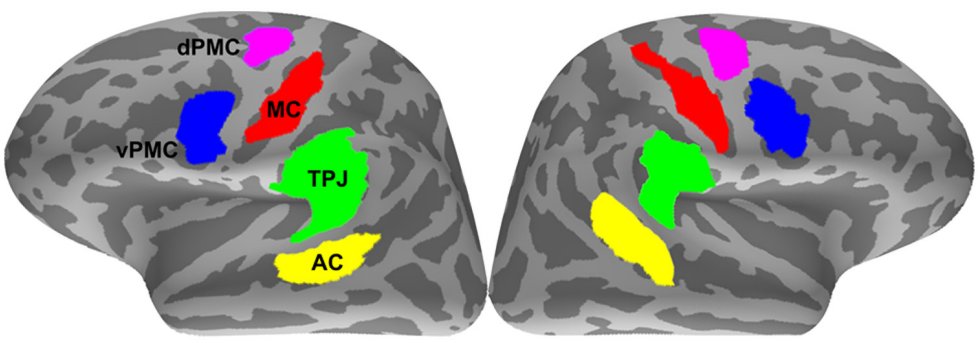


FIGURE 2 | Regions-of-interest (ROIs). AC, auditory cortex; TPJ, temporoparietal junction; MC, motor cortex; vPMC, ventral premotor cortex; dPMC, dorsal premotor cortex.

responses) a surrogate data was created by adopting a trial shuffle approach (Lachaux et al., 1999). One thousand artificial trial orders were generated by randomly shuffling the trials in each ROI independently. For each randomization, WPLIs were calculated as described in Section "Phase Synchrony Estimation". A p-value was acquired by determining the percentage of the surrogate values exceeding the original WPLI (or correlation coefficient in the correlation tests). The null hypothesis (i.e., phase synchrony results are explained by the regions independently synchronizing to the stimulus onset) was rejected at $p < 0.05$.

For estimating directionality of information flow for the significant functional connections, a *post hoc* SEM analysis was conducted (Penny et al., 2004). The SEM was performed in the same time and frequency range as the given phase synchrony effect. Continuous wavelet transform was applied to decompose the ROI time courses into time-frequency representations, similarly to the phase synchrony calculations. As samples in MEG time series are not independent, which can lead to inflated correlation between ROIs and thus bias the estimated path coefficients, the significance of the estimated paths was quantified with a bootstrap approach allowing the statistical inferences on the estimated paths to be based on empirical, rather than theoretical, estimates of the null distribution of path coefficients.

Pairwise path coefficients were tested for models with reciprocal connections between ROIs (i.e., ROI1→ROI2→ROI1). Statistical significance was tested across subjects with a paired-samples permutation t-test on the path coefficients (β) of the directed connections (i.e., $\beta_{A\rightarrow B}$ vs. $\beta_{B\rightarrow A}$). The goodness-of-fit between the model and data was tested with the root mean square error of approximation (RMSEA; Steiger, 1990), based on the chi-square test statistic (Pearson, 1900). A RMSEA value less than 0.07 is considered a good fit (Steiger, 2007).

All analyses and statistical tests on phase synchrony were implemented in Python, with the help of MNE-Python (Gramfort et al., 2014) and SciPy toolkit (http://www.scipy.org/). Analyses and statistical tests on SEM were implemented in MATLAB (Mathworks, Natick, MA, USA) using custom scripts and computer resources within the Aalto Science-IT project.

RESULTS

BEHAVIORAL RESULTS

Phonetic categorization performance was quantified as the ratio of correctly vs. incorrectly identified noisy syllables in the active listening condition involving overt repetition (/pa/ vs. /ta/; mean d-prime = 1.29, SD = 0.95; mean percent correct = 70.4%, for /pa/ 62.4%, for /ta/ 78.0%, SD = 13.6%).

INTER-AREAL NEURAL SYNCHRONY

Effect of stimulus type and condition

Figure 3 shows the effects of intelligibility (noisy vs. clear stimuli) and task (active vs. passive listening) as well as their interaction on inter-areal neural synchrony. Only the significant time-frequency points that coincided with significant values as compared to the trial-shuffled null distribution are reported.

Stronger neural synchrony was observed in response to noisy compared to intact syllables between two pairs of left-hemisphere ROIs: (1) AC and vPMC from 60–80 ms ∼23 Hz [$F(1,19) = 36.5$, pFDR = 0.008]; and (2) dPMC and TPJ from 190–200 ms at ∼23–26 Hz [$F(1,19) = 34.9$, pFDR = 0.02; **Figure 3A**]. The intact stimuli did not elicit stronger neural synchrony than the noisy stimuli between any pairs of ROIs.

Stronger neural synchrony was found in active compared to passive listening condition for (1) left TPJ and vPMC from 120–130 ms at ∼38 Hz [$F(1,19) = 27.1$, pFDR = 0.04]; and (2) right TPJ and vPMC from 170–200 ms at ∼71–74 Hz [$F(1,19) = 43.3$, pFDR = 0.001; **Figure 3B**]. None of the ROI pairs showed stronger synchrony in passive compared to active listening condition.

Significant condition x stimulus type interaction was observed between left AC and vPMC from 60–80 ms ∼20–23 Hz [$F(1,19) = 44.6$, pFDR = 0.0008]. *Post hoc* t-test revealed that this was caused by stronger synchrony in response to noisy speech only in the passive listening condition (**Figure 3C**). All F- and p-values are from the time-frequency point of strongest effect.

Direction of information flow between the ROI pairs that showed significant synchrony effects was assessed using the pairwise path coefficients obtained with SEM (depicted with arrows in **Figure 3**). Directed interactions were found from left AC to vPMC [$t(19) = 8.14$, $p < 0.001$], from left dPMC to TPJ [$t(19) = 2.78$, $p = 0.02$], and from left vPMC to TPJ [$t(19) = 3.02$, $p = 0.01$]. No significant directionality was found between the right vPMC and TPJ [$t(19) = 0.93$, $p = 0.36$].

Correlation with speech sound identification accuracy

As shown in **Figure 4**, speech sound identification accuracy correlated positively with four left-hemisphere connections: (1) between AC and TPJ from 60–80 ms after stimulus onset at ∼23 Hz (spearman $r = 0.83$, pFDR = 0.002); (2) between AC and vPMC from 90–110 ms at ∼20–23 Hz (spearman $r = 0.80$, pFDR = 0.006); (3) between TPJ and vPMC from 90–120 ms at ∼17–23 Hz (spearman $r = 0.76$, pFDR = 0.02), and (4) between vPMC and MC from 120–140 ms at ∼11–14 Hz (spearman $r = 0.74$, pFDR = 0.03). The correlation coefficients and p-values are from the time-frequency point of strongest correlation. Correlation between phase synchrony and syllable identification accuracy was not found with respect to the left dPMC or between any right-hemispheric ROIs.

The trial-shuffling analysis showed that all the phase synchrony effects remained significant after controlling for the possibility that the ROIs were independently synchronizing to the stimulus onset. The p-values (averaged across the significant time-frequency points) for the significance of the residual induced phase synchrony were as follows: AC–TPJ ($p = 0.001$), AC–vPMC ($p = 0.007$), TPJ–vPMC ($p = 0.007$), and vPMC–MC ($p = 0.003$). The speech sound identification performance showed no statistical outliers or correlation with subjects' age (spearman $r = -0.09$, $p = 0.69$; age range 21–58 years, with one subject aged over 40), diminishing the possibility that the findings could be explained by age-related audiological and brain differences.

To estimate the direction of information flow, pairwise path coefficients obtained with SEM were tested (depicted with arrows in **Figure 4**). Directed interactions were found from AC to TPJ [$t(19) = 8.30$, $p < 0.001$], from AC to vPMC [$t(19) = 2.36$, $p = 0.03$], and from vPMC to TPJ [$t(19) = 2.42$, $p = 0.03$].

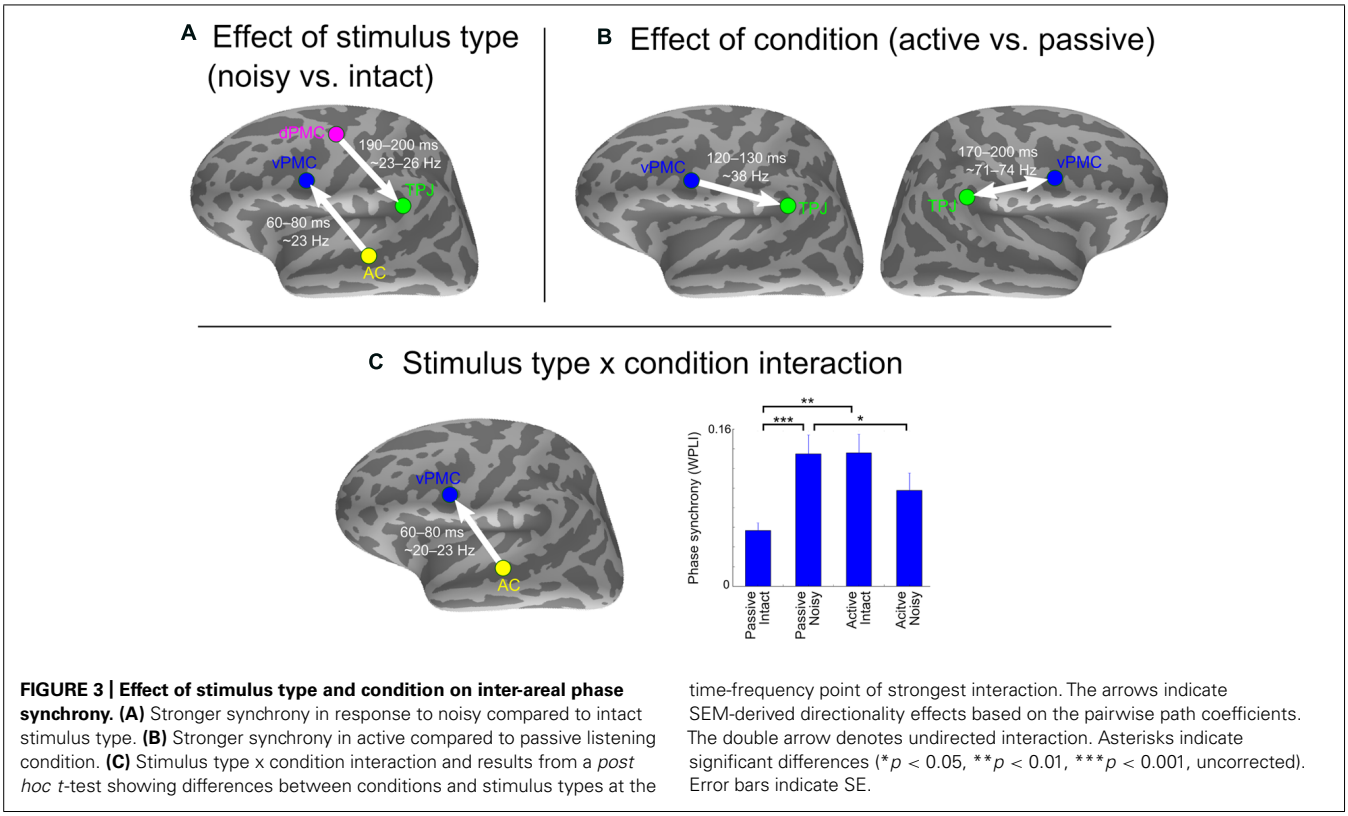


FIGURE 3 | Effect of stimulus type and condition on inter-areal phase synchrony. (A) Stronger synchrony in response to noisy compared to intact stimulus type. **(B)** Stronger synchrony in active compared to passive listening condition. **(C)** Stimulus type x condition interaction and results from a *post hoc t*-test showing differences between conditions and stimulus types at the time-frequency point of strongest interaction. The arrows indicate SEM-derived directionality effects based on the pairwise path coefficients. The double arrow denotes undirected interaction. Asterisks indicate significant differences ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$, uncorrected). Error bars indicate SE.

No significant directionality was found between vPMC and MC [$t(19) = 0.23$, $p = 0.81$].

Finally, as shown in **Figure 5**, model comparison was performed between the three functionally interconnected left-hemisphere areas (i.e., AC, TPJ, and vPMC) to determine the model of information flow that best fits the data within the 50–200 ms time window. To avoid the possible bias introduced by comparing models with different degrees of freedom, only unidirectional connections were defined, resulting in a total of 8 candidate models. Two models exhibited mean RMSEA smaller than 0.07, indicating a good fit to the data (Steiger, 2007): AC→vPMC→TPJ→AC (RMSEA: 0.058 ± 0.024; mean ± SD) and AC→TPJ→vPMC→AC (RMSEA: 0.062 ± 0.025; mean ± SD).

DISCUSSION

The present study examined inter-areal synchrony of neuronal oscillations during speech perception. MEG was recorded while subjects were (1) passively listening to auditory speech sounds (/pa/ and /ta/) presented with or without acoustic noise and (2) engaged in a sensorimotor task involving the identification and overt repetition of the same sounds.

Synchrony between four pairs of left-hemisphere regions showed positive correlation with speech sound identification accuracy within 150 ms after stimulus onset (**Figure 4**). The correlation between AC and TPJ occurred at ~23 Hz and peaked early (60–80 ms). This was followed by correlations between AC and vPMC (90–110 ms at ~20 Hz), TPJ and vPMC (90–120 ms at ~17–23 Hz), and lastly between vPMC and MC (120–140 ms at ~11–14 Hz). *Post hoc* analysis with SEM suggested that information flows from AC to TPJ, from AC to vPMC, and from vPMC to TPJ (**Figure 4**).

These findings suggest that neural communication between auditory speech processing areas and motor cortical areas facilitates phonetic categorization and that the left TPJ functions as an interface where auditory signals are matched with articulatory-motor information. The directed interaction from AC to vPMC and from vPMC to TPJ could be reflecting a processing loop whereby the acoustic speech activates articulatory-motor representations and generates a forward prediction containing information of the sensory consequences of realizing those motor commands. The directed interaction from AC to TPJ between 60–80 ms, on the other hand, could be reflecting a quick sketch of the sensory event (Bar et al., 2006), which is compared against the forward prediction (Rauschecker, 2011). The sensory expectation generated by the forward prediction would then serve to complement the acoustic information for improved phonetic categorization. The SEM model comparison supports the existence of such sensorimotor loops, indicating that models where information flow between the left AC, TPJ, and vPMC forms a loop in either direction fits well to the data (**Figure 5**). This interpretation is in line with the "perception-for-action-control theory" (PACT; Schwartz et al., 2012), according to which speech percepts are shaped by both sensory processing and motor knowledge of speech gestures. As phonetic categorization performance was quantified in the active listening condition involving overt repetition, inherent task differences need to be considered when interpreting the results. However, since access to

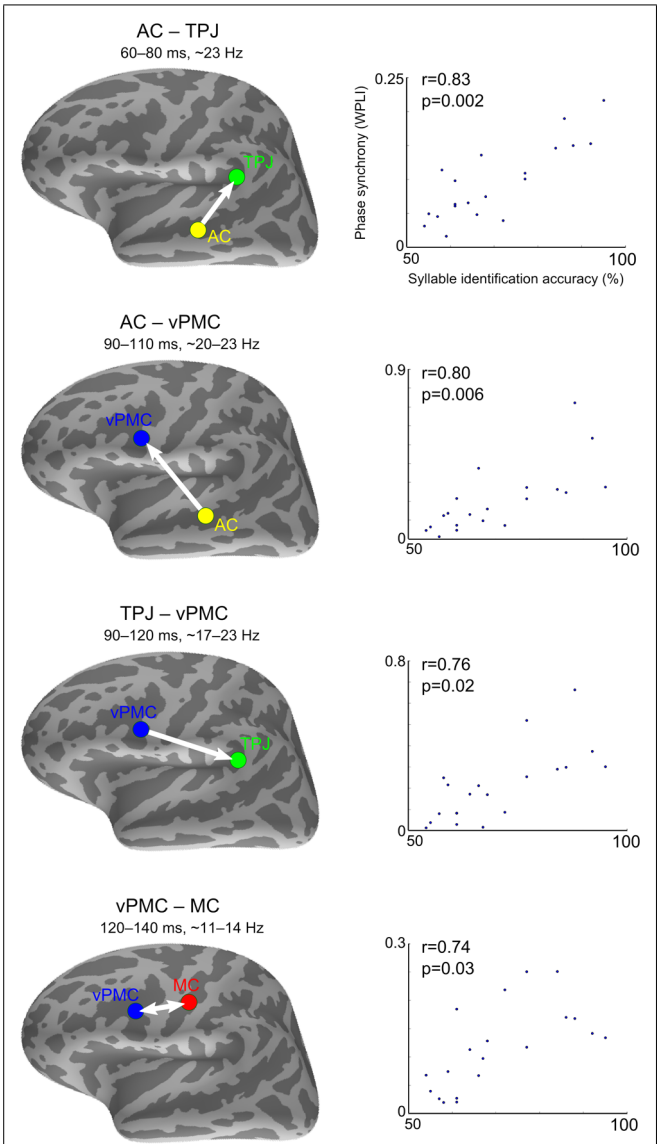


FIGURE 4 | Correlations between inter-areal phase synchrony and syllable identification accuracy. Syllable identification scores plotted against phase synchrony strength (WPLI) at the time-frequency point of strongest correlation. The spearman rank correlation coefficients (*r*) and corresponding *p*-values are denoted in each plot. The arrows indicate SEM-derived directionality effects based on the pairwise path coefficients. The double arrow denotes undirected interaction. AC, auditory cortex; TPJ, temporoparietal junction; vPMC, ventral premotor cortex; MC, motor cortex.

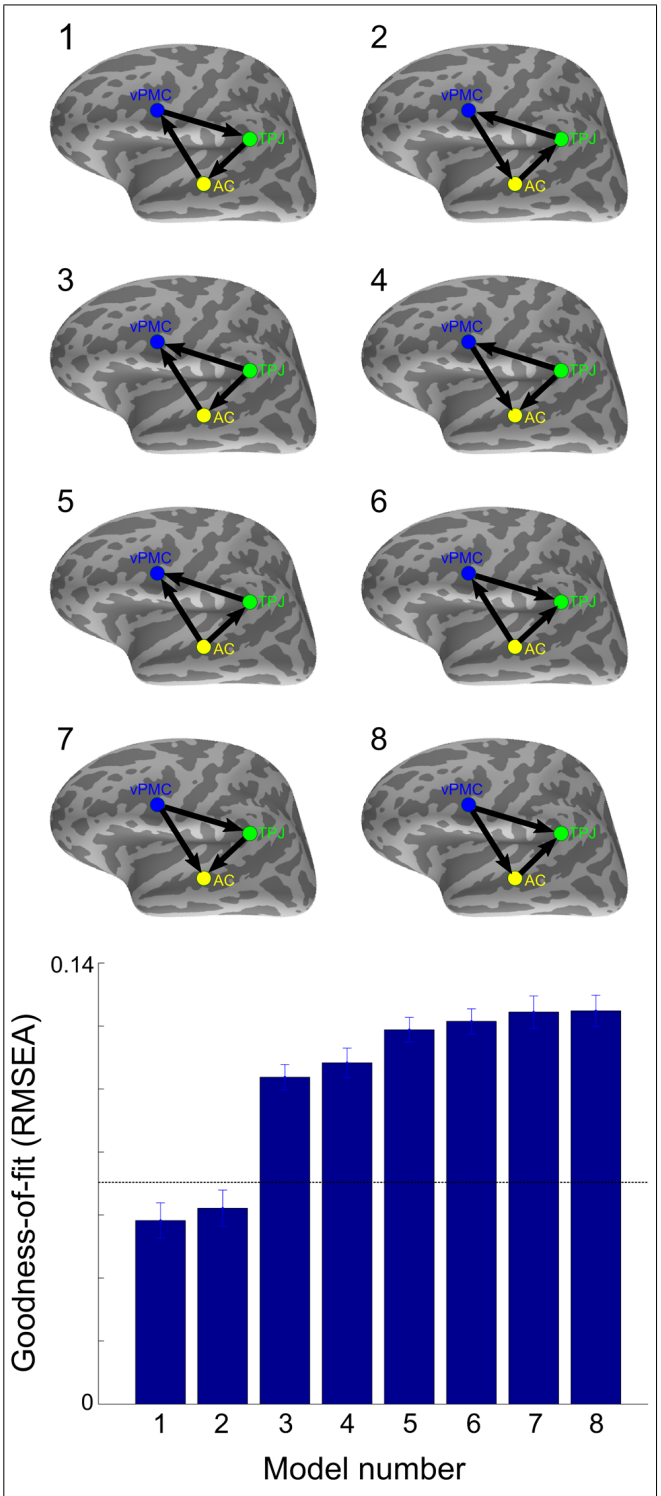


FIGURE 5 | Comparison between models of effective connectivity. RMSEA was applied to test the goodness-of-fit between all unidirectional SEM models between the three functionally interconnected left-hemisphere areas. The horizontal dashed line denotes the cut-off point with RMSEA < 0.07 considered a good fit (Steiger, 2007). Error bars indicate SE. AC, auditory cortex; TPJ, temporoparietal junction; vPMC, ventral premotor cortex.

phonological categories occurs at ~150 ms after sound onset (for a review, see Salmelin, 2007) and since the observed effects occurred within 150 ms after sound onset, it is unlikely that they are reflecting speech preparation (e.g., mental rehearsal) while waiting for the appearance of the visual cue to overtly repeat.

The present results are consistent with our earlier study (Alho et al., 2012), in which positive correlation was found between syllable identification accuracy and PMC response amplitudes at

∼100 ms after stimulus onset. These results, along with the findings of another recent study (Szenkovits et al., 2012), suggest also that PMC recruitment varies across subjects, which can be due to individual differences in, e.g., phonological short-term memory (Seghier and Price, 2009). Consistently, Chevillet et al. (2013) found that subjects with more category-selective PMC representations (as observed using fMRI rapid adaptation paradigm) were better able to categorize phonemes in a behavioral test after scanning, thus implying that the representation might be recruited to assist explicit phonetic categorization. The observed individual differences in speech sound identification accuracy can also be explained by differences in allocation of attention. It is noteworthy, however, that selective attention and forward prediction in sensorimotor integration might be supported by similar neural mechanisms. Indeed, the sensory expectation generated by the forward prediction can be understood as increased gain for processing, or reshaping of neuronal receptive fields to be more selective to, the attended/expected auditory features (Hickok et al., 2011). The mechanism for sensorimotor integration could thus, similarly to selective attention, induce short-term plasticity effects on the AC (for a review, see Jääskeläinen and Ahveninen, 2014), and therefore enhance behavioral performance, such as sound discrimination (Kauramäki et al., 2007; Ahveninen et al., 2011). Relatedly, a recent study demonstrated, by using TMS and MEG, that when speech sounds were attended, the articulatory-motor cortex contributed to the auditory processing of the sounds already at 60–100 ms after sound onset, whereas when unattended, the contributing effect started considerably later, at ∼170 ms after sound onset (Möttönen et al., 2014). These findings suggest that, although the motor contribution to speech processing seems to occur automatically (Chevillet et al., 2013; Möttönen et al., 2013), early sensorimotor interactions are dependent on attention.

Notably, the phase synchrony effects among AC, TPJ, and vPMC occurred in the beta frequency band (∼20 Hz), which is compatible with previous studies revealing an association between beta-band synchrony and sensorimotor integration (for a review, see Siegel et al., 2012). Furthermore, as successful speech perception requires temporal integration of information with high modulation frequency (e.g., formant transitions in /pa/ vs. /ta/), it can be argued that the brain oscillations involved in such a cognitive process must correspond to this frequency (Giraud and Poeppel, 2012). Beta-band oscillations could therefore be sufficiently rapid for the coordination among anatomically distributed neuronal assemblies during encoding and integration of speech information.

Complementing the correlational findings, ANOVA showed a main effect of intelligibility (i.e., noisy vs. clear speech) with stronger synchrony first between left AC and vPMC and later between left TPJ and dPMC for noisy compared to clear speech. Such increase in neural synchrony between auditory and motor regions appears compatible with previous fMRI studies showing a stronger recruitment of motor regions in case of ambiguous stimuli, as e.g., during masked or distorted vs. intelligible speech or during auditory identification of non-native vs. native phonemes (Binder et al., 2004; Callan et al., 2004; Wilson and Iacoboni, 2006; Zekveld et al., 2006). This finding, together with the strong intelligibility x task interaction between left AC and vPMC (caused by enhanced synchrony for noisy compared to clear stimuli only during passive listening; **Figure 3C**) suggests that frontal motor areas support the sensory processing of degraded speech automatically, in the absence of tasks or explicit attention directed to the speech sounds (although, see Wild et al., 2012). As information flow was estimated from AC to vPMC and from dPMC to TPJ, the results converge with findings demonstrating a mediating effect of top-down feedback in the disambiguation of speech (e.g., Gow and Segawa, 2009). The main effect of task (i.e., active vs. passive listening) provided evidence for stronger synchrony between TPJ and vPMC in both hemispheres during active compared to passive perception task, which is likely reflecting enhanced sensorimotor integration (i.e., mapping between auditory and articulatory-motor representations) when people are actively engaged in a speech decision task with subsequent oral responses. This finding is concordant with a recent study showing bilateral sensorimotor transformations during perception in an overt speech repetition task (Cogan et al., 2014) and another showing that while passive listening to speech involved only temporal areas, active speech comprehension was recruiting also bilateral inferior frontal areas (Yue et al., 2013). The left-hemisphere connection showed directionality from vPMC to TPJ, possibly reflecting the integration of motor knowledge with speech inputs through top-down predictions (or attentional modulation, as previously discussed).

In conclusion, our results showed that (1) engagement in a sensorimotor task involving speech sound identification and overt repetition enhanced connectivity bilaterally between the TPJ and vPMC within 200 ms after sound onset; (2) passive listening to noisy speech elicited stronger connectivity than clear speech between left AC and vPMC at ∼100 ms, and between left dPMC and TPJ at ∼200 ms; and (3) connectivity strength among left AC, vPMC, and TPJ correlated positively with speech sound identification accuracy. The estimated directions of information flow support the idea that top-down feedback from the articulatory-motor areas influences low-level phonetic processing. Taken together, these findings suggest that sensorimotor integration mediates the categorization of incoming speech sounds through reciprocal auditory-to-motor and motor-to-auditory projections.

ACKNOWLEDGMENTS

This study was financially supported by the Academy of Finland (projects 257811, 130412, and 138145), and by research grants from CNRS (Center National de la Recherche Scientifique) and ANR (Agence Nationale de la Recherche, ANR SPIM and MULTI-STAP) to Marc Sato. The authors declare no competing financial interests.

REFERENCES

Ahveninen, J., Hämäläinen, M., Jääskeläinen, I. P., Ahlfors, S. P., Huang, S., Lin, F. H., et al. (2011). Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4182–4187. doi: 10.1073/pnas.1016134108

Alho, J., Sato, M., Sams, M., Schwartz, J. L., Tiitinen, H., and Jääskeläinen, I. P. (2012). Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. *Neuroimage* 60, 1937–1946. doi: 10.1016/j.neuroimage.2012.02.011

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci. U.S.A.* 103, 449–454. doi: 10.1073/pnas.0507062103

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125, 279–284. doi: 10.1016/S0166-4328(01)00297-2

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., and Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nat. Neurosci.* 7, 295–301. doi: 10.1038/nn1198

Callan, D. E., Jones, J. A., Callan, A. M., and Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *Neuroimage* 22, 1182–1194. doi: 10.1016/j.neuroimage.2004.03.006

Cappa, S. F., and Pulvermuller, F. (2012). Cortex special issue: language and the motor system. *Cortex* 48, 785–787. doi: 10.1016/j.cortex.2012.04.010

Chevillet, M. A., Jiang, X., Rauschecker, J. P., and Riesenhuber, M. (2013). Automatic phoneme category selectivity in the dorsal auditory stream. *J. Neurosci.* 33, 5208–5215. doi: 10.1523/JNEUROSCI.1870-12.2013

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., and Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. *Nature* 507, 94–98. doi: 10.1038/nature12935

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2011). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex* 48, 882–887. doi: 10.1016/j.cortex.2011.05.017

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., et al. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* 26, 55–67. doi: 10.1016/S0896-6273(00)81138-1

Davis, M. H., and Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *J. Neurosci.* 23, 3423–3431.

Davis, M. H., and Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hear. Res.* 229, 132–147. doi: 10.1016/j.heares.2007.01.014

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010

Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* 8, 272–284. doi: 10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4

Giraud, A. L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063

Gow, D. W. Jr., and Segawa, J. A. (2009). Articulatory mediation of speech perception: a causal analysis of multi-modal imaging data. *Cognition* 110, 222–236. doi: 10.1016/j.cognition.2008.11.011

Grabski, K., Tremblay, P., Gracco, V. L., Girin, L., and Sato, M. (2013). A mediating role of the auditory dorsal pathway in selective adaptation to speech: a state-dependent transcranial magnetic stimulation study. *Brain Res.* 1515, 55–65. doi: 10.1016/j.brainres.2013.03.024

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Hämäläinen, M. S., and Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Med. Biol. Eng. Comput.* 32, 35–42. doi: 10.1007/bf02512476

Hämäläinen, M. S., and Sarvas, J. (1989). Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. *IEEE Trans. Biomed. Eng.* 36, 165–171. doi: 10.1109/10.16463

Hickok, G. (2012). The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. *J. Commun. Disord.* 45, 393–402. doi: 10.1016/j.jcomdis.2012.06.004

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hipp, J. F., Engel, A. K., and Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron* 69, 387–396. doi: 10.1016/j.neuron.2010.12.027

Huang, S., Chang, W. T., Belliveau, J. W., Hamalainen, M., and Ahveninen, J. (2014). Lateralized parietotemporal oscillatory phase synchronization during auditory selective attention. *Neuroimage* 86, 461–469. doi: 10.1016/j.neuroimage.2013.10.043

Jääskeläinen, I. P., and Ahveninen, J. (2014). Auditory-cortex short-term plasticity induced by selective attention. *Neural Plast.* 2014, 11. doi: 10.1155/2014/216731

Kauramäki, J., Jääskeläinen, I. P., and Sams, M. (2007). Selective attention increases both gain and feature selectivity of the human auditory cortex. *PLoS ONE* 2:e909. doi: 10.1371/journal.pone.0000909

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

Kujala, J., Pammer, K., Cornelissen, P., Roebroeck, A., Formisano, E., and Salmelin, R. (2007). Phase coupling in a cerebro-cerebellar network at 8–13 Hz during reading. *Cereb. Cortex* 17, 1476–1485. doi: 10.1093/cercor/bhl059

Kveraga, K., Ghuman, A. S., Kassam, K. S., Aminoff, E. A., Hämäläinen, M. S., Chaumon, M., et al. (2011). Early onset of neural synchronization in the contextual associations network. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3389–3394. doi: 10.1073/pnas.1013760108

Lachaux, J. P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Hum. Brain Mapp.* 8, 194–208. doi: 10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lin, F. H., Belliveau, J. W., Dale, A. M., and Hämäläinen, M. S. (2006). Distributed current estimates using cortical orientation constraints. *Hum. Brain Mapp.* 27, 1–13. doi: 10.1002/hbm.20155

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Möttönen, R., Dutton, R., and Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cereb. Cortex* 23, 1190–1197. doi: 10.1093/cercor/bhs110

Möttönen, R., Van De Ven, G. M., and Watkins, K. E. (2014). Attention fine-tunes auditory-motor processing of speech sounds. *J. Neurosci.* 34, 4064–4069. doi: 10.1523/JNEUROSCI.2214-13.2014

Möttönen, R., and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825. doi: 10.1523/JNEUROSCI.6018-08.2009

Palva, J. M., Monto, S., Kulashekhar, S., and Palva, S. (2010). Neuronal synchrony reveals working memory networks and predicts individual memory capacity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 7580–7585. doi: 10.1073/pnas.0913113107

Palva, S., and Palva, J. M. (2012). Discovering oscillatory interaction networks with M/EEG: challenges and breakthroughs. *Trends Cogn. Sci.* 16, 219–230. doi: 10.1016/j.tics.2012.02.004

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser.* 5 50, 157–175. doi: 10.1080/14786440009463897

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004). Modelling functional integration: a comparison of structural equation and dynamic causal models. *Neuroimage* 23(Suppl. 1), S264–S274. doi: 10.1016/j.neuroimage.2004.07.041

Rauschecker, J. P. (2011). An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hear. Res.* 271, 16–25. doi: 10.1016/j.heares.2010.09.001

Rauschecker, J. P., and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Salmelin, R. (2007). Clinical neurophysiology of language: the MEG approach. *Clin. Neurophysiol.* 118, 237–254. doi: 10.1016/j.clinph.2006.07.316

Sato, M., Tremblay, P., and Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7. doi: 10.1016/j.bandl.2009.03.002

Schwartz, J.-L., Basirat, A., Ménard, L., and Sato, M. (2012). The perception-for-action-control theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguistics* 25, 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Seghier, M. L., and Price, C. J. (2009). Dissociating functional brain networks by decoding the between-subject variability. *Neuroimage* 45, 349–359. doi: 10.1016/j.neuroimage.2008.12.017

Siegel, M., Donner, T. H., and Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nat. Rev. Neurosci.* 13, 121–134. doi: 10.1038/nrn3137

Singer, W. (2009). Distributed processing and temporal codes in neuronal networks. *Cogn. Neurodyn.* 3, 189–196. doi: 10.1007/s11571-009-9087-z

Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *J. Neurosci.* 32, 8443–8453. doi: 10.1523/JNEUROSCI.5069-11.2012

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502_4

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Pers. Individ. Dif.* 42, 893–898. doi: 10.1016/j.paid.2006.09.017

Szenkovits, G., Peelle, J. E., Norris, D., and Davis, M. H. (2012). Individual differences in premotor and motor recruitment during speech perception. *Neuropsychologia* 50, 1380–1392. doi: 10.1016/j.neuropsychologia.2012.02.023

Vinck, M., Oostenveld, R., Van Wingerden, M., Battaglia, F., and Pennartz, C. M. (2011). An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *Neuroimage* 55, 1548–1565. doi: 10.1016/j.neuroimage.2011.01.055

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012). Effortful listening: the processing of degraded speech depends critically on attention. *J. Neurosci.* 32, 14010–14021. doi: 10.1523/JNEUROSCI.1528-12.2012

Wilson, S. M., and Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage* 33, 316–325. doi: 10.1016/j.neuroimage.2006.05.032

Womelsdorf, T., Schoffelen, J. M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., et al. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science* 316, 1609–1612. doi: 10.1126/science.1139597

Yue, Q., Zhang, L., Xu, G., Shu, H., and Li, P. (2013). Task-modulated activation and functional connectivity of the temporal and frontal areas during speech comprehension. *Neuroscience* 237, 87–95. doi: 10.1016/j.neuroscience.2012.12.067

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., and Schoonhoven, R. (2006). Top-down and bottom-up processes in speech comprehension. *Neuroimage* 32, 1826–1836. doi: 10.1016/j.neuroimage.2006.04.199

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 January 2014; accepted: 14 April 2014; published online: 06 May 2014.
Citation: Alho J, Lin F-H, Sato M, Tiitinen H, Sams M and Jääskeläinen IP (2014) Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. Front. Psychol. ***5****:394. doi: 10.3389/fpsyg.2014.00394*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 July 2014
doi: 10.3389/fpsyg.2014.00727

Effect of attentional load on audiovisual speech perception: evidence from ERPs

Agnès Alsius[1], Riikka Möttönen[2], Mikko E. Sams[3], Salvador Soto-Faraco[4,5] and Kaisa Tiippana[6]*

[1] Psychology Department, Queen's University, Kingston, ON, Canada
[2] Department of Experimental Psychology, University of Oxford, Oxford, UK
[3] Brain and Mind Laboratory, School of Science, Aalto University, Espoo, Finland
[4] Institut Català de Recerca i Estudis Avançats, Barcelona, Spain
[5] Brain and Cognition Center, Universitat Pompeu Fabra, Barcelona, Spain
[6] Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

Edited by:
Jean-Luc Schwartz, CNRS, France

Reviewed by:
Michael Pilling, Oxford Brookes University, UK
Sophie Bouton, University of Geneva, Switzerland

***Correspondence:**
Kaisa Tiippana, Division of Cognitive and Neuropsychology, Institute of Behavioural Sciences, University of Helsinki, PO Box 9, Helsinki 00014, Finland
e-mail: kaisa.tiippana@helsinki.fi

Seeing articulatory movements influences perception of auditory speech. This is often reflected in a shortened latency of auditory event-related potentials (ERPs) generated in the auditory cortex. The present study addressed whether this early neural correlate of audiovisual interaction is modulated by attention. We recorded ERPs in 15 subjects while they were presented with auditory, visual, and audiovisual spoken syllables. Audiovisual stimuli consisted of incongruent auditory and visual components known to elicit a McGurk effect, i.e., a visually driven alteration in the auditory speech percept. In a Dual task condition, participants were asked to identify spoken syllables whilst monitoring a rapid visual stream of pictures for targets, i.e., they had to divide their attention. In a Single task condition, participants identified the syllables without any other tasks, i.e., they were asked to ignore the pictures and focus their attention fully on the spoken syllables. The McGurk effect was weaker in the Dual task than in the Single task condition, indicating an effect of attentional load on audiovisual speech perception. Early auditory ERP components, N1 and P2, peaked earlier to audiovisual stimuli than to auditory stimuli when attention was fully focused on syllables, indicating neurophysiological audiovisual interaction. This latency decrement was reduced when attention was loaded, suggesting that attention influences early neural processing of audiovisual speech. We conclude that reduced attention weakens the interaction between vision and audition in speech.

Keywords: audiovisual speech perception, multisensory integration, McGurk effect, attention, event-related potentials

INTRODUCTION

Many events in our everyday life stimulate different sensory systems in a correlated fashion. The integration of such diversity of sensory information allows the brain to construct efficient and adaptive representations of the external world (e.g., Stein and Meredith, 1993), but the neural mechanisms underlying multisensory binding are still not well understood (e.g., van Atteveldt et al., 2014). A question under current debate is to which extent multisensory integration occurs pre-attentively or can be influenced by higher-order cognitive processes (e.g., Talsma et al., 2010).

Speech perception is one of the classical examples of multisensory binding in humans, whereby acoustic information is combined with the sight of corresponding facial articulatory gestures. Audiovisual association of facial gestures and vocal sounds has been demonstrated in non-human primates (Ghazanfar and Logothetis, 2003) and in pre-linguistic children (e.g., Kuhl and Meltzoff, 1982; Burnham and Dodd, 2004; Pons et al., 2009), arguing for the existence of an early basis of this capacity (Soto-Faraco et al., 2012). One striking demonstration of multisensory binding in speech is the McGurk effect (McGurk and MacDonald, 1976), which results from exposure to mismatched acoustic and visual signals, often leading observers to hear an illusory speech sound. For example, when the sound of [ba] is dubbed onto a video clip containing the articulatory movements corresponding to [ga], the observer usually experiences hearing a fusion between the acoustic and the visual syllable, e.g., [da] or [tha], or even the visually specified [ga]. Discrepant visual speech thus alters the auditory speech percept, and may even dominate it, e.g., a visual [da] dubbed onto an acoustic [ba] is often heard as [da], and a visual [na] dubbed onto an acoustic [ma] is heard as [na] (MacDonald and McGurk, 1978; for a detailed discussion on the definition of the McGurk effect, see Tiippana, 2014). The compelling phenomenology of the McGurk illusion has been often used as an argument supporting the effortless and mandatory (i.e., unavoidable) nature of multisensory integration in speech (e.g., Rosenblum and Saldaña, 1996; Soto-Faraco et al., 2004).

Several recent studies have, however, put into question the impenetrability of audiovisual integration to attentional modulation, both in the speech (Tiippana et al., 2004, 2011; Alsius et al., 2005, 2007; Soto-Faraco and Alsius, 2007, 2009; Andersen et al., 2009; Fairhall and Macaluso, 2009; Alsius and Soto-Faraco, 2011;

Buchan and Munhall, 2011, 2012) and the non-speech domains (e.g., Senkowski et al., 2005; Talsma and Woldorff, 2005; Fujisaki et al., 2006; Talsma et al., 2007). Of particular interest for the current study, Alsius et al. (2005) tested to which extent audiovisual speech perception could be modulated by attentional load. They varied the amount of available processing resources by measuring the participants' susceptibility to the McGurk effect in a Single vs. Dual task paradigm. In the Dual task condition, participants were instructed to perform a very demanding detection task on rapidly presented visual or auditory streams, while repeating back the words uttered by a speaker (which were dubbed to obtain the McGurk effect). In the Single task condition, participants were shown the same displays but just prompted to repeat back the words. In the Dual task condition, the percentage of illusory McGurk responses decreased dramatically. That is, when the load was high, and thus processing resources presumably depleted, participants became less susceptible to experience the McGurk effect than when they had spare processing resources.

Effects of attention on multisensory processing have been reported also outside the domain of speech, for example using event-related potentials (ERPs). Talsma and Woldorff (2005; see also Senkowski et al., 2005; Talsma et al., 2007) showed that the difference usually found between the evoked potentials to audiovisual (AV) events and the sum of unisensory events (A+V; "additive model") was larger at attended than unattended locations of space. This modulation was seen both in short and long latency ERP components. Talsma et al.'s (2007) study suggests that spatial attention affects the early sensory integration of simple (non-speech) multisensory events. It remains unknown, however, how attentional load (as in Alsius et al., 2005) modulates the neural mechanisms underlying audiovisual speech integration.

Electrophysiological studies within the speech domain have consistently shown that visual speech can modify activity in the auditory cortex during audiovisual speech perception as early as ∼100–200 ms post-stimulus (Sams et al., 1991; Colin et al., 2002; Möttönen et al., 2002, 2004; Klucharev et al., 2003; Besle et al., 2004; Van Wassenhove et al., 2005). There are a variety of electrophysiogical markers of audiovisual interactions in speech (e.g., Saint-Amour et al., 2007; Bernstein et al., 2008; Ponton et al., 2009; Arnal et al., 2011). Although these markers are not exclusive of audiovisual speech (Stekelenburg and Vroomen, 2007), they are thought to reflect important aspects of the speech perception process such as cross-modal prediction and phonological processing (Brunellière et al., 2013).

One of the best-known electrophysiological correlates of audiovisual interactions in speech is temporal facilitation of the N1/P2 component of the auditory ERPs (Van Wassenhove et al., 2005; Baart et al., 2014; Knowland et al., 2014). Some studies have also found an amplitude reduction of the N1/P2 complex in audiovisual speech contexts (Klucharev et al., 2003; Besle et al., 2004; Van Wassenhove et al., 2005; Pilling, 2009; Knowland et al., 2014), but this effect has not always been replicated (Miki et al., 2004; Möttönen et al., 2004; Baart et al., 2014). It is also relevant here to note that studies on the effect of attention on the auditory evoked potentials have often focused on modulations within the N1 and P2 time windows, generally demonstrating an amplification of these ERP components when the stimulus is under the focus of attention (see Hillyard et al., 1973; Picton et al., 1974; Näätänen, 1982 for seminal studies).

The goal of the present study was to characterize the role of attentional load in audiovisual integration of speech, capitalizing on the electrophysiological marker of temporal facilitation. The amount of processing resources directed to audiovisual stimuli was manipulated by using a Single vs. Dual task paradigm adapted from Alsius et al. (2005, 2007). ERPs were recorded while participants were presented with audiovisual spoken syllables known to produce the McGurk effect, as well as unisensory auditory and visual syllables. These were interspersed within an Rapid Serial Visual Presentation (RSVP) of line drawings. In the Single task condition, participants were asked to identify some of the syllables regardless of the RSVP, whereas in the Dual task condition, participants were asked to perform the syllable identification task and, in addition, to detect repetitions in the RSVP.

We expected that audiovisual interaction would modulate the N1/P2 component complex of the auditory ERPs in the Single task condition, as shown in previous studies (e.g., Van Wassenhove et al., 2005; Baart et al., 2014; Knowland et al., 2014). Crucially, with respect to the attentional load, we hypothesized that these modulations would be reduced or eliminated in the Dual task condition if early audiovisual interactions in the auditory cortex are influenced by attention demands. We thus predicted that the temporal facilitation of the N1/P2 complex for audiovisual ERPs would be smaller in the Dual than Single task condition.

METHODS

PARTICIPANTS

Sixteen healthy right-handed participants, native speakers of Finnish, participated in the experiment. Data from two participants were excluded from the analyses because of excessive artifacts in EEG signals. In the remaining 14 participants, the mean age was 22 years (range 19–28 years; 3 female). Participants reported normal audition and normal or corrected-to-normal vision. All of them gave their informed consent to participate in the study. The study was conducted in accordance with the principles expressed in the Declaration of Helsinki, and adhered to the guidelines of the American Psychological Society and the ethical policies of Helsinki University of Technology (currently Aalto University; please note that at the time of data collection, there was no ethical committee at the university from which to apply for approval).

STIMULI

Digital video recordings of a Finnish female speaker (black-and-white, full-face) uttering the syllables [mi] and [ni] were edited with Studio Purple software and transformed to bitmap sequences. The image contrast was lowered to minimize visual ERP responses. The auditory components of the syllables were saved as 16 Bit—44.1 kHz waveform audio format (WAV) files. The auditory unisensory trials consisted of an acoustic syllable [mi] or [ni] combined with a still image of the talker's face with the lips closed. The visual unisensory trials consisted of the silent presentation of the speaker's articulation of the [mi] or [ni] syllable (presented as a sequence of still images, 25 frames

per second). The McGurk-type audiovisual trials were created by temporally aligning the acoustic burst onset of the auditory syllable [mi] to the burst onset of the visual [ni]. This particular combination is known to elicit an auditory percept dominated by the visual information so that observers usually hear /ni/ (MacDonald and McGurk, 1978; Tiippana et al., 2011, where the same stimuli were used as here). Each visual syllable was presented in a clip of 600 ms duration (15 frames), and each auditory syllable lasted 265 ms. In the audiovisual stimuli, the auditory syllable started 215 ms after the onset of visual articulatory gestures (5th frame).

There were two experimental conditions run in different blocks (Single task and Dual task condition, see Procedure). Each block contained a sequence of a total of 180 audiovisual (AV) syllables presented in random order (120 McGurk stimuli, 30 congruent [mi], 30 congruent [ni]), 150 visual-only (V) syllables (120 [ni], 30 [mi]), and 150 auditory-alone (A) syllables (120 [mi], 30 [ni]). The inter-syllable interval was chosen randomly between 1200 and 3600 ms (in order to minimize anticipatory slow waves) contained a still picture of the talkers face. After ∼10% of the syllables (a total of 10 times per stimulus-type, in each condition) and distributed randomly in the sequence, the question "What did you hear?" appeared on the screen, prompting participants to make an identification response on the last syllable presented. The syllable sequence was interspersed within a RSVP stream of line drawings of common objects presented in between syllables (3–6 drawings at each inter-syllable period), and superimposed on the still image of the talker's face. The RVSP stopped while syllables were presented in order to prevent overlapping ERPs to pictures and syllables. Nevertheless, monitoring had to be sustained across these breaks because repetitions could straddle syllable presentations.

In the RSVP, each drawing was presented for 60 ms, stimulus onset asynchrony (SOA) varied randomly between 400 and 600 ms, and they roughly covered the distance between the upper part of the speaker's lips and the nose. Each drawing in the sequence was chosen at random from a set of 105 different drawings from the Snodgrass and Vanderwart (1980) picture database, and rotated in one of three possible different orientations (45, 90, or 135°, equiprobably). Picture repetitions (i.e., targets in the Dual task condition) occurred on average every seven stimuli, and could occur within or across the inter-syllable period.

The stimulus presentation protocol was controlled using Presentation software (Neurobehavioural system, Inc.). Images were presented using a 19″ CRT monitor. Sounds were delivered at an overall intensity of 65 dB(A) SPL through two loudspeakers positioned on both sides of the monitor.

PROCEDURE

Participants sat 1.1 m from the monitor on a comfortable armchair placed in an electrically and acoustically shielded room. They were instructed to make a syllable identification response when prompted to (on ∼10% of the trials) by pressing the corresponding key on the keyboard (labeled "mi" or "ni"). The amount of available processing resources directed to the spoken syllables was manipulated by the instructions regarding a concurrent task. Whereas in the Single task condition participants just had to identify the syllable when prompted, in the Dual task condition participants were asked to, in addition to the identification response, continuously monitor the RSVP of line drawings superimposed on the image of the talker's face for repetitions, and respond by pressing a key labeled "X" on the keyboard when repetitions occurred (see **Figure 1**). All participants were tested in both the Dual and the Single task condition. The order of the tasks

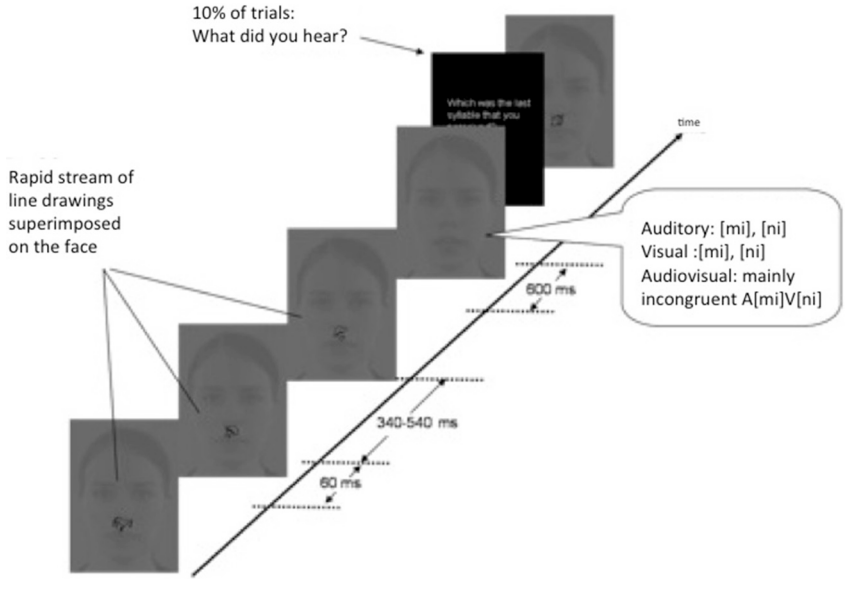


FIGURE 1 | Setup. A Single vs. Dual task paradigm was used to divert attention from spoken syllables. In the Single task condition participants reported the syllables that the speaker said (10% of trials), whereas in the Dual task they were asked, in addition, to monitor a Rapid Visual Serial Presentation stream of line drawings for repeated items. The syllables were presented audiovisually, auditorily, and visually.

was counterbalanced between participants. A training block was performed before starting each task.

EEG DATA ACQUISITION

EEG recordings were made using BrainVision software with 20 silver/silver chloride electrodes (BrainCap, Brainproducts) mounted on an elastic cap (reduced 10–20 system: Fp1, Fp2, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, Fz, Cz, Pz, TP9, TP10). Two additional electro-oculogram electrodes (Eog1 and Eog2) were placed above and below eyes in order to detect blink artifacts, and one electrode was attached to the tip of the nose in order to provide a common reference channel. Prior to each session, all electrode impedances were set below 10 kΩ. EEG data were recorded with a sampling frequency of 500 Hz.

DATA ANALYSIS

ERPs were averaged offline separately for the three stimulus types (auditory [mi], visual [ni], and audiovisual McGurk stimulus A[mi]+V[ni]) using Vision Analyzer software. The tip of the nose was selected as the reference for the analysis. Data were filtered using a bandpass of 1–40 Hz (attenuation 24 dB/octave) and segmented in time windows of −100 to 400 ms relative to the auditory onset of the syllable (i.e., the zero time corresponding to the onset of the sound, or the onset of the 5th video frame for the visual-only trials). A 100-ms pre-stimulus (before the auditory onset) baseline was used. Trials with signal amplitudes exceeding 100 μV at any electrode within the −100 to +400 ms window were automatically rejected to eliminate response contamination by eye movements or muscular activities. Trials in which a motor response was produced to any of the two tasks at any time between 100 ms prior to 400 ms after the syllable was presented were also excluded from the ERP analyses. The averaged ERPs for each subject and condition contained a minimum of 100 epochs after trial rejection. In order to ensure sufficient number of observations, the EEG session was extended when the number of artifacts detected during the experiment was high.

Estimation of AV interactions

AV interactions were assessed by using a modified version of a commonly used additive model: AV-[A+V] (Stein and Meredith, 1993; Giard and Peronnet, 1999; Molholm et al., 2002; Teder-Sälejärvi et al., 2002; Klucharev et al., 2003; Besle et al., 2004; Möttönen et al., 2004). As we specifically focused on the modulation of auditory ERPs, which have been shown to be prominent during audiovisual speech processing, we compared the ERPs evoked by the unisensory auditory stimulus (A) with the subtraction between the ERPs evoked by the audiovisual (AV) and visual (V) stimuli, i.e., AV-V (Baart et al., 2014). The AV-V wave represents the EEG activity evoked by the audiovisual syllables without the contribution of the visual component. Differences between the AV-V wave and the A wave should reveal how audiovisual interaction affects N1 and P2 in Single and Dual task conditions.

The A and AV-V waveforms were statistically compared by performing sample-by-sample (∼2 ms steps) sequential paired Student *t*-tests and by comparing the peak latencies and amplitudes of the N1 and P2 components of the auditory ERPs in both Single and Dual task conditions. The sample-by-sample student *t*-tests were performed from audio onset to 300 ms post-audio onset in all electrodes for the data from the Single and the Dual task conditions. In order to reduce the likelihood of false-positives due to a large number of *t*-tests, we considered differences to be significant when the *p* values were lower than 0.05 at 10 (=20 ms) or more consecutive time points (Guthrie and Buchwald, 1991; see also Molholm et al., 2002; Besle et al., 2004 for the same analysis procedure).

The Fz electrode was selected for comparison of A and AV-V. Electrode selection was necessary since in many electrodes the RSVP elicited more pronounced and longer-lasting ERPs during the Dual than Single task condition, which could contaminate the baselines to the speech stimuli. In the Fz recording site, the baseline was not contaminated, the N1 and P2 responses to A stimuli were the strongest, and the differences between A and AV-V were maximal.

The N1 peak was defined as the largest negative peak occurring between 65 and 165 ms after the auditory onset at Fz from A and AV-V ERPs. The P2 peak was computed as the highest positive value in a temporal window of 135–285 ms after the onset of the auditory stimulus. After semi-automatic detection of the peaks, two experimenters blind to the subject's condition visually revised that each detected peak had been correctly identified.

RESULTS

BEHAVIORAL RESULTS

Syllable identification

For each stimulus type (AV, V, A) we assessed the proportion of visually-influenced responses. The data were submitted to repeated measures ANOVA with two within-participants factors: Stimulus type (AV, V, A) and Task (Single, Dual). The main effects of Task and Stimulus type were both significant [$F_{(1, 13)} = 23.49$, $p < 0.001$; $F_{(2, 26)} = 20.11$, $p < 0.001$, respectively] and so was the interaction between them [$F_{(2, 26)} = 8.85$, $p = 0.001$]. When each stimulus type was analyzed separately, significant effect of the Task was observed for both AV and V stimuli ($t = 4.1$, $p = 0.001$ and $t = 4.4$, $p = 0.001$, respectively), but it did not affect the identification of A stimuli ($t = 0.00$, $p = 1$). That is, the percentage of participants' visually-influenced responses was significantly lower in the Dual than Single task condition for audiovisual and visual stimuli. No difference was found in the size of this decrease between AV and V [$F_{(1, 13)} = 0.18$, $p = 0.68$; **Figure 2**]. These results mean that the McGurk effect was weaker and speechreading poorer in the Dual than Single task condition.

Target detection in the concurrent task of the Dual task condition

In the concurrent repetition task (Dual task condition), the overall hit rate (detection response within 2 s after a target occurred in the RSVP stream) was 0.35 (note that the average probability of target occurrence was 1 every 7), and false alarm rate (erroneously responding when no target occurred within the previous 2 s) was 0.008.

ELECTROPHYSIOLOGICAL RESULTS

Audiovisual interactions in the single task condition

Figure 3A shows the grand-average ERPs to the A stimuli and the AV-V difference wave at Fz in the Single task condition. In

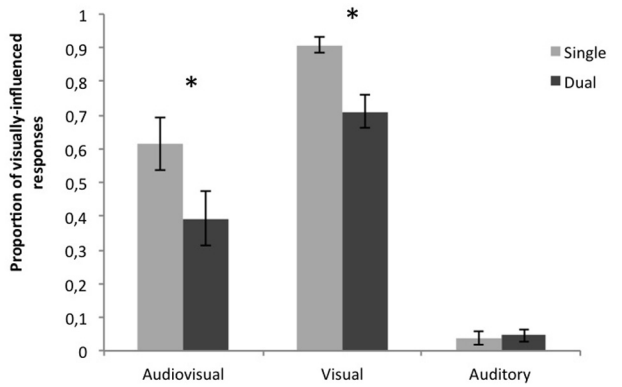


FIGURE 2 | Proportions of visually-influenced responses in the behavioral task for audiovisual, visual and auditory stimuli in Single and Dual task conditions. For audiovisual, and visual stimuli, the proportion of visually-influenced responses (i.e., correct responses to V stimuli and "ni" responses to AV stimuli) was significantly reduced in the Dual task condition. The asterisk denotes significant differences ($p < 0.05$).

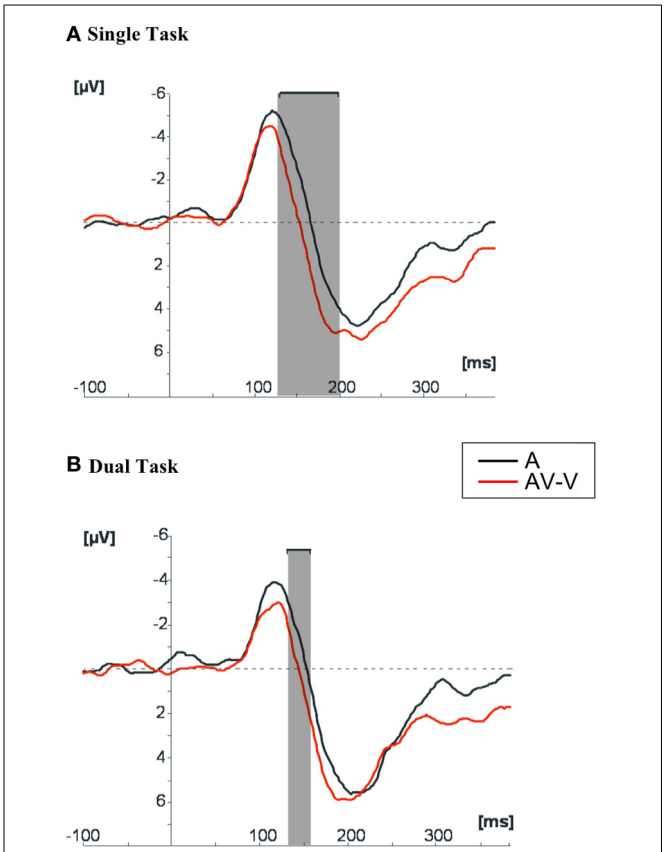


FIGURE 3 | Grand average of the ERPs to auditory stimuli in comparison with the ERPs resulting from the subtraction AV-V at Fz, in (A) Single task condition and (B) Dual task condition. The shaded areas indicate the time windows in which the two ERPs differed significantly ($p < 0.05$) in amplitude.

the early time window (100–140 ms) both responses were characterized by the typical negative N1 component originating in the auditory cortex (Vaughan and Ritter, 1970; Picton et al., 1974). N1 was followed by a P2 component. Paired sequential *t*-tests showed a reliable difference between AV-V and A ERPs from 130 to 200 ms (all $p < 0.05$) after the auditory onset. This was because of the earlier occurrence of the N1 offset and P2 onset in the AV-V wave than in A, suggesting that auditory responses were speeded up by the presentation of concurrent visual speech information (see peak latency analysis below).

The topographical distribution of the ERPs to the A, AV-V, and (AV-V)-A difference wave (**Figure 4**) support the assumption that the difference between A and AV-V ERPs was due to modulation of auditory ERPs. In the ERPs to A stimuli, N1 peaked at 122 ms and was maximal at fronto-central sites (Fz: −5.670 μV) with a polarity inversion at the mastoids (TP9: 0.659 μV; TP10: 0.649 μV). The auditory P2 peaked at 221 ms at Fz (5.76 μV) with a polarity inversion at the mastoids (TP9: −0.79 μV; TP10: −0.30 μV). These distributions of ERPs to acoustic stimuli can be attributed to dipolar current sources in the auditory cortex (Vaughan and Ritter, 1970; Scherg and Von Cramon, 1986). The distributions of AV-V ERPs resembled those of the ERPs to unisensory A stimuli, suggesting similar neural generators. That is, N1 peaked at 114 ms and was maximal at Fz (−4.99 μV) with the minimal negativity observed at mastoids (TP9: −0.377 μV; TP10: −0.512 μV) and P2 peaked at 204 ms at Fz (6.12 μV) and showed reversed polarity at mastoids (TP9: −1.25 μV; TP10: −0.17 μV).

Importantly, the scalp distribution of the (AV-V)-A difference (see time points 160 and 190 ms in **Figure 4**) was similar to that of the P2 response to A stimuli (see time points 190 and 220 ms in **Figure 4**). The difference (AV-V)-A was also maximal at fronto-central scalp sites with polarity inversion at the mastoids (see time points 160 and 190 in **Figure 4**). Thus, the cerebral sources of the interaction term (AV-V)-A are likely to be similar to the ones of the auditory ERPs, suggesting that the neural generators of auditory ERPs in the auditory cortices were modulated by audiovisual interaction.

Effect of processing load on audiovisual interactions (Single vs. Dual task conditions)

Figure 3B shows the grand average ERPs at Fz obtained to the presentation of auditory stimuli and the AV-V difference wave in the Dual task condition. The difference between A and AV-V in the Dual task condition was significant during a short 20-ms time window (135–155 ms), compared to the 70-ms time window (130–200 ms) in the Single task condition. The difference between A and AV-V in Single and Dual tasks could not be attributed to amplitude differences, since repeated measures ANOVAs for the peak amplitudes of N1 and P2 with Modality (A, AV-V) and Task (Single, Dual) as factors showed no significant main effects or interactions.

In order to further test whether visual speech speeded up auditory processing in Single and Dual task conditions, we performed separate repeated measures ANOVAs for the peak latencies of N1 and P2 with Modality (A, AV-V) and Task (Single, Dual) as factors. Because we wanted to test a directional hypothesis that temporal facilitation should be smaller in the Dual than Single

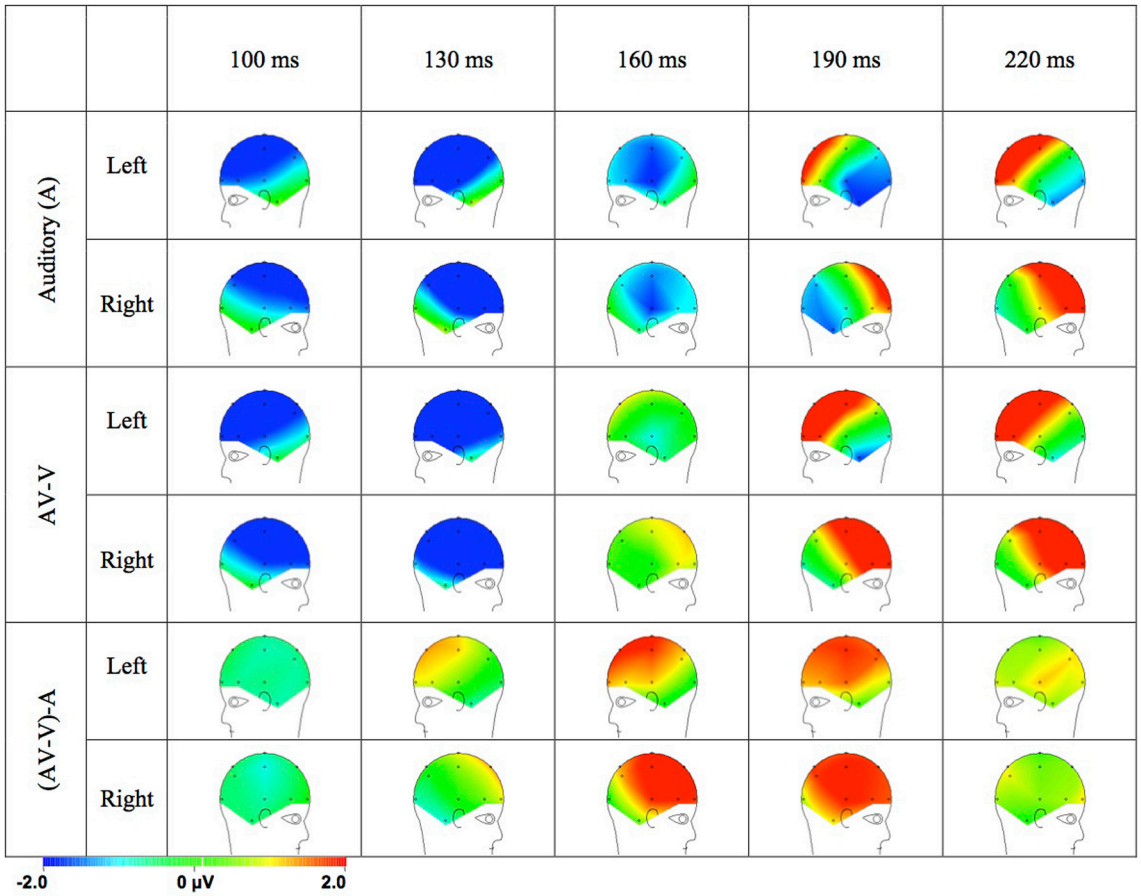


FIGURE 4 | Topographical distributions of the grand average ERPs for the auditory stimuli and AV-V and the (AV-V)-A difference waves in time steps of 30 ms.

task condition, we also carried out planned comparisons (t-tests) on the contrast A>(AV-V) in Dual and Single task conditions for both N1 and P2.

The main effect of Modality was significant for both N1 [$F_{(1,\ 13)} = 5.92$, $p < 0.05$] and P2 [$F_{(1,\ 13)} = 7.01$, $p < 0.05$], but the main effect of Task was not [N1: $F_{(1,\ 13)} = 0.229$, $p = 0.64$; P2: $F_{(1,\ 13)} = 3.67$, $p = 0.08$], nor was the interaction [N1: $F_{(1,\ 13)} = 1.96$, $p = 0.184$; P2: $F_{(1,\ 13)} = 1.25$, $p = 0.0284$]. The main effect of Modality arose because the latencies were overall shorter in AV-V than A for both task conditions (Dual, Single) and ERP components (N1, P2) (**Figure 5**).

The planned comparisons, testing the hypothesis that temporal facilitation decreased when processing resources are loaded, showed that the latency shifts between A and AV-V modalities were statistically significant only in the Single Task condition. That is, N1 peaked earlier in AV-V than in A [114 and 122 ms, respectively; $t_{(13)} = 2.34$, $p < 0.05$] in the Single Task condition, whereas in the Dual task condition the latency shift of N1 was not significant [115 and 117 ms, respectively; $t_{(13)} = 0.804$, $p = 0.436$]. In a similar fashion, P2 peaked significantly earlier in AV-V than in A [204 and 221 ms, respectively; $t_{(13)} = 2.34$, $p < 0.05$] in the Single task condition, but P2 latency shift was not significant in the Dual task condition [197 and 205 ms, respectively; $t_{(13)} = 1.67$, $p = 0.118$]. That is, when participants focused attention on a difficult unrelated visual task, the temporal facilitatory effects on the N1/P2 complex tended to be reduced or to disappear. Probably, the fact that in all cases, the AV-V latency peaks were numerically shorter than the A peaks prevented the interaction term of the ANOVA to reach significance between Task and Modality, a tendency that was nevertheless captured by the individual t-tests. Thus, these results are well in line with the predicted effect of attention on AV speech processing, but the conclusions (based on the t-tests) must be qualified by the fact that the overall ANOVAs did not reveal significant interactions.

DISCUSSION

To evaluate the role of attention in audiovisual speech perception, we measured behavioral and electrophysiological responses to audiovisual, auditory and visual speech stimuli in a Single vs. Dual task paradigm. Results from both measures converged to the idea that increasing demands on visual attentional resources exerted a detrimental effect on the outcome of multisensory speech processing.

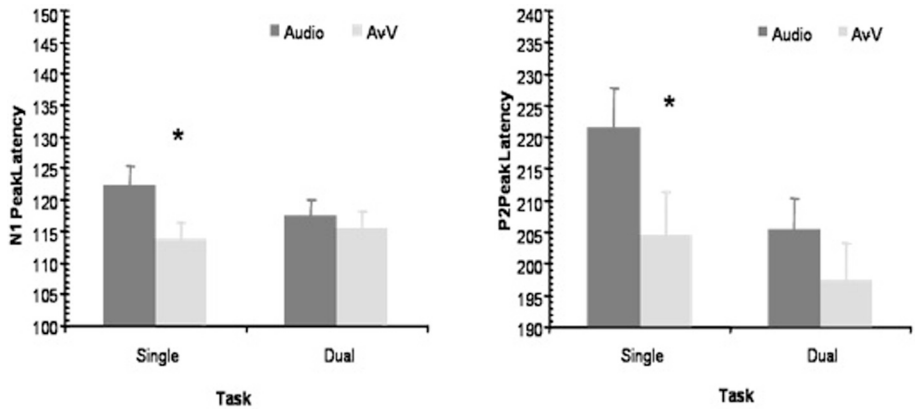


FIGURE 5 | Peak latency modulations. The peak latencies of N1 and P2 at Fz were significantly reduced in the AV-V signal in comparison to those evoked by auditory stimuli in the Single task condition, but not in the Dual task condition. The asterisk denotes significant differences between A and AV-V ($p < 0.05$).

The behavioral results showed that the McGurk effect was weaker in the Dual than Single task condition, showing an attentional effect on audiovisual speech perception, in agreement with previous results (Tiippana et al., 2004, 2011; Alsius et al., 2005, 2007; Soto-Faraco and Alsius, 2007, 2009; Andersen et al., 2009; Alsius and Soto-Faraco, 2011; Buchan and Munhall, 2011, 2012). However, note that at variance with the results of Alsius et al. (2005; see also Alsius et al., 2007), the identification of visual stimuli was poorer in the Dual than Single task condition. Thus, the attention effect in this study could in principle be attributed to a modulation exerted by visual attention on a modality-specific stage, interfering with the processing of visual speech prior to multisensory integration (Massaro, 1998; Tiippana et al., 2004). This interpretation has to be put under the light of electrophysiological and other recent evidence highlighting the flexible nature of the interplay between multisensory integration and attention. Indeed, there is a variety of possible stages and mechanisms enabling multisensory integration and, therefore, the impact of attention in integration processes might express in different ways (Talsma et al., 2010; van Atteveldt et al., 2014).

Our electrophysiological results replicated the previous finding (Van Wassenhove et al., 2005; Baart et al., 2014; Knowland et al., 2014) that the latency of the N1/P2 complex is reduced for audiovisual compared to auditory speech stimuli. This suggests that the visual component of audiovisual speech speeds up processing of the acoustic input, possibly in the auditory cortex (Van Wassenhove et al., 2005). When comparing peak latencies in the Single and Dual task conditions, the AV-V signal peaked significantly earlier than the A signal in the Single task condition, in which the processing resources could be fully devoted to audiovisual stimuli. Yet, when participants' processing resources were diverted to a concurrent visual task in the Dual task condition, the latency difference between the AV-V and A ERPs was non-significant. It should be noted, though, that no significant interaction between Modality and Task was found. This lack of interaction is likely to be due to the presence of some integration effect in both Single and Dual task conditions, and it advises for some caution in the interpretation of the results. Yet, what is clear is that, when tested for the specific prediction that the temporal facilitation for audiovisual ERPs would be smaller in the Dual than Single task condition, the prediction was confirmed since the facilitation was significant in the Single, but not in the Dual task condition. Supporting this conclusion, the window of significant differences between AV-V and A in the sample by sample analyses was larger in the Single Task condition (70 ms) than in the Dual Task condition (20 ms).

The electrophysiological temporal facilitation was beyond any unisensory effect since in the model used here (A vs. AV-V), any attentional effects on visual processing should have been canceled out when subtracting the visual ERPs from the audiovisual ERPs, and therefore can be ruled out as a cause of the differences. Based on the polarity and scalp topography of the difference (AV-V)-A—which was maximally positive over the fronto-central regions of the scalp and inverted in polarity in the mastoids—it is likely that the audiovisual interaction effect stems from modulation of auditory processing. This interaction, observed in the Single task condition and found to be sensitive to attentional load in the Dual task condition, was likely to be generated in the auditory cortices. The current ERP evidence thus lends some support to the view that taxing processing resources may interfere with multisensory interactions in the auditory cortex to some extent.

In absolute terms, the latency values were highest for auditory stimuli in the Single task condition. However, we think that the safest way to interpret the present pattern of results is in relative terms, not in absolute ones. This is because the baseline modulation produced by attention onto each modality separately might not be the same. Therefore, the focus should be on how AV-V peak latencies change with respect to the "default" A latency, within each attention condition. This comparison revealed a decrease in the Single, but not in the Dual task condition.

From a functional perspective, our results are in keeping with the notion that during speech perception, the auditory and visual sensory systems interact at multiple levels of processing (Schwartz et al., 1998; Nahorna et al., 2012; Barrós-Loscertales et al., 2013), and that top-down modulatory signals can influence at least some of these levels. Multisensory links do not solely rely

on feed-forward convergence from unisensory regions to multisensory brain areas, but also implicate back-projections from association areas to multiple levels of (early) sensory processing that are based on current task demands (Calvert et al., 1999, 2000; Macaluso et al., 2000; Friston, 2005; Driver and Noesselt, 2008). This kind of recurrent architecture naturally allows for an integral role of attention during multisensory integration (Driver and Spence, 2000; Frith and Driver, 2000; Talsma et al., 2010; van Atteveldt et al., 2014).

Given the current evidence, briefly sketched above, we argue that since attention can influence processing at multiple levels, visual attentional load can interfere with unisensory visual processing involved in speechreading, resulting in poorer identification of visual speech, as well as with multisensory integration even at early processing stages, resulting in reduced temporal facilitation of auditory evoked potentials by audiovisual speech.

In conclusion, the present results provide new insights into the cognitive and neural mechanisms underlying audiovisual speech integration, as they suggest that visual processing load can modulate early stages of audiovisual processing. Our findings further challenge the view that audiovisual speech integration proceeds in a strictly bottom-up sensory-driven manner, independently of attention.

ACKNOWLEDGMENTS

This research was supported by grants from Spanish *Ministry of Science and Innovation* (PSI2010-15426) and the European Research Council (StG-2010 263145) to Salvador Soto-Faraco and Agnès Alsius; and by the Academy of Finland. Agnès Alsius was supported by a BRD scholarship from the University of Barcelona. Kaisa Tiippana, Mikko E. Sams, and Riikka Möttönen were supported by the Academy of Finland. Correspondence concerning this article should be addressed to Kaisa Tiippana (email: kaisa.tiippana@helsinki.fi).

REFERENCES

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., and Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Exp. Brain Res.* 183, 399–404. doi: 10.1007/s00221-007-1110-1

Alsius, A., and Soto-Faraco, S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. *Exp. Brain Res.* 213, 175–183. doi: 10.1007/s00221-011-2624-0

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., and Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Commun.* 51, 184–193. doi: 10.1016/j.specom.2008.07.004

Arnal, L. H., Wyart, V., and Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. Neurosci.* 14, 797–801. doi: 10.1038/nn.2810

Baart, M., Stekelenburg, J. J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 53, 115–121. doi: 10.1016/j.neuropsychologia.2013.11.011

Barrós-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Ávila Rivera, C., et al. (2013). Neural correlates of audiovisual speech processing in a second language. *Brain Lang.* 126, 253–262. doi: 10.1016/j.bandl.2013.05.009

Bernstein, L. E., Auer, E. T. Jr., Wagner, M., and Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. *Neuroimage* 39, 423–435. doi: 10.1016/j.neuroimage.2007.08.035

Besle, J., Fort, A., Delpuech, C., and Giard, M. H. (2004). Bimodal speech: early suppressive effects in the human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Brunellière, A., Sánchez-García, C., Ikumi, N., and Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *Int. J. Psychophysiol.* 89, 136–147. doi: 10.1016/j.ijpsycho.2013.06.016

Buchan, J. N., and Munhall, K. G. (2011). The influence of selective attention to auditory and visual speech on the integration of audiovisual speech information. *Perception* 40, 1164–1182. doi: 10.1068/p6939

Buchan, J. N., and Munhall, K. G. (2012). The effect of a concurrent working memory task and temporal offsets on the integration of auditory and visual speech information. *Seeing Perceiving* 25, 87–106. doi: 10.1163/187847611X620937

Burnham, D., and Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Dev. Psychobiol.* 45, 204–220. doi: 10.1002/dev.20032

Calvert, G. A., Brammer, M. J., Bullmure, E. T., Campbell, R., Iversen, S. D., and David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10, 2619–2623. doi: 10.1097/00001756-199908200-00033

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10, 649–657. doi: 10.1016/S0960-9822(00)00513-3

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., and Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin. Neurophysiol.* 113, 495–506. doi: 10.1016/S1388-2457(02)00024-X

Driver, J., and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. *Neuron* 57, 11–23. doi: 10.1016/j.neuron.2007.12.013

Driver, J., and Spence, C. (2000). Multisensory perception: Beyond modularity and converge. *Curr. Biol.* 10, R731–R735. doi: 10.1016/S0960-9822(00)00740-5

Fairhall, S. L., and Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29, 1247–1257. doi: 10.1111/j.1460-9568.2009.06688.x

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Frith, C., and Driver, J. (2000). Shifting baselines in attention research. *Nat. Rev. Neurosci.* 1, 147–148. doi: 10.1038/35039083

Fujisaki, W., Koene, A., Arnold, D., Johnston, A., and Nishida, S. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proc. R. Soc. B Biol. Sci.* 273, 865–874. doi: 10.1098/rspb.2005.3327

Ghazanfar, A. A., and Logothetis, N. K. (2003). Facial expressions linked to monkey calls. *Nature* 423, 937–938. doi: 10.1038/423937a

Giard, M., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 1, 473–490. doi: 10.1162/089892999563544

Guthrie, D., and Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology* 28, 240–244. doi: 10.1111/j.1469-8986.1991.tb00417.x

Hillyard, S. A., Hink, R. F., Schwent, V. L., and Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science* 182, 177–180. doi: 10.1126/science.182.4108.177

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Knowland, V. C., Mercure, E., Karmiloff-Smith, A., Dick, F., and Thomas, M. S. (2014). Audio-visual speech perception: a developmental ERP investigation. *Dev. Sci.* 17, 110–124. doi: 10.1111/desc.12098

Kuhl, P. K., and Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science* 218, 1138–1141. doi: 10.1126/science.7146899

Macaluso, E., Frith, C. D., and Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science* 289, 1206–1208. doi: 10.1126/science.289.5482.1206

MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. *Percept. Psychophys.* 24, 253–257. doi: 10.3758/BF03206096

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* Cambridge, MA: MIT Press.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 265, 746–748. doi: 10.1038/264746a0

Miki, K., Watanabe, S., and Kakigi, R. (2004). Interaction between auditory and visual stimulus relating to the vowel sounds in the auditory cortex in humans: a magnetoencephalographic study. *Neurosci. Lett.* 357, 199–202. doi: 10.1016/j.neulet.2003.12.082

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cogn. Brain Res.* 14, 115–128. doi: 10.1016/S0926-6410(02)00066-6

Möttönen, R., Krause, C. M., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cogn. Brain Res.* 13, 417–425. doi: 10.1016/S0926-6410(02)00053-8

Möttönen, R., Schürmann, M., and Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neurosci. Lett.* 363, 112–115. doi: 10.1016/j.neulet.2004.03.076

Näätänen, R. (1982). Processing negativity: an evoked-potential reflection of selective attention. *Psychol. Bull.* 92, 605–640. doi: 10.1037/0033-2909.92.3.605

Nahorna, O., Berthommier, F., and Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* 132, 1061–1077. doi: 10.1121/1.4728187

Picton, T. W., Hillyard, S. A., Krausz, H. I., and Galambos, R. (1974). Human auditory evoked potentials. I: evaluation of components. *Electroencephalogr. Clin. Neurophysiol.* 36, 179–190. doi: 10.1016/0013-4694(74)90155-2

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J. Speech Lang. Hear. Res.* 52, 1073–1081. doi: 10.1044/1092-4388(2009/07-0276)

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., and Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proc. Natl. Acad. Sci.* 106, 10598–10602. doi: 10.1073/pnas.0904134106

Ponton, C. W., Bernstein, L. E., and Auer, E. T., Jr. (2009). Mismatch negativity with visual-only and audiovisual speech. *Brain Topogr.* 21, 207–215. doi: 10.1007/s10548-009-0094-5

Rosenblum, L. D., and Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 22, 318–331.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., and Foxe, J. J. (2007). Seeing voices: high-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia* 45, 587–597. doi: 10.1016/j.neuropsychologia.2006.03.036

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-F

Scherg, M., and Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalogr. Clin. Neurophysiol.* 65, 344–360. doi: 10.1016/0168-5597(86)90014-6

Schwartz, J.-L., Robert-Ribes, J., and Escudier, P. (1998). "Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, eds R. Campbell, B. Dodd, and D. Burnham (Hove: Psychology Press), 85–108.

Senkowski, D., Talsma, D., Herrmann, C. S., and Woldorff, M. G. (2005). Multisensory processing and oscillatory gamma responses: effects of spatial selective attention. *Exp. Brain Res.* 166, 411–426. doi: 10.1007/s00221-005-2381-z

Snodgrass, J. G., and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Hum. Learn. Mem.* 6, 174–215. doi: 10.1037/0278-7393.6.2.174

Soto-Faraco, S., and Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport* 18, 347–350. doi: 10.1097/WNR.0b013e32801776f9

Soto-Faraco, S., and Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 580–587. doi: 10.1037/a0013483

Soto-Faraco, S., Calabresi, M., Navarra, J., Werker, J., and Lewkowicz, D. J. (2012). The development of audiovisual speech perception. *Multisensory Dev.* 207–228. doi: 10.1093/acprof:oso/9780199586059.003.0009

Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* 92, B13–B23. doi: 10.1016/j.cognition.2003.10.005

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, MA: MIT Press.

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.* 19, 1964–1973. doi: 10.1162/jocn.2007.19.12.1964

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Talsma, D., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 7, 1098–1114. doi: 10.1162/0898929054475172

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., and Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event–related potentials. *Cogn. Brain Res.* 14, 106–114. doi: 10.1016/S0926-6410(02)00065-4

Tiippana, K. (2014). What is the McGurk Effect? *Front. Psychol.* 5:725. doi: 10.3389/fpsyg.2014.00725

Tiippana, K., Andersen, T. S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

Tiippana, K., Puharinen, H., Möttönen, R., and Sams, M. (2011). Sound location can influence audiovisual speech perception when spatial attention is manipulated. *Seeing Perceiving* 24, 67–90. doi: 10.1163/187847511X557308

van Atteveldt, N., Murray, M. M., Thut, G., and Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron* 81, 1240–1253. doi: 10.1016/j.neuron.2014.02.044

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

Vaughan, H. G. Jr., and Ritter, W. (1970). The sources of auditory evoked responses recorded from the human scalp. *Electroencephalogr. Clin. Neurophysiol.* 28, 360–367. doi: 10.1016/0013-4694(70)90228-2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2014; accepted: 23 June 2014; published online: 15 July 2014.
*Citation: Alsius A, Möttönen R, Sams ME, Soto-Faraco S and Tiippana K (2014) Effect of attentional load on audiovisual speech perception: evidence from ERPs. Front. Psychol. **5**:727. doi: 10.3389/fpsyg.2014.00727*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 01 July 2014
doi: 10.3389/fpsyg.2014.00678

Hearing impairment and audiovisual speech integration ability: a case study report

Nicholas Altieri* and Daniel Hudock

Department of Communication Sciences and Disorders, Idaho State University, Pocatello, ID, USA

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Annalisa Setti, University College Cork, Ireland
Jacqueline Leybaert, Université Libre de Bruxelles, Belgium

****Correspondence:***
Nicholas Altieri, Department of Communication Sciences and Disorders, Idaho State University, 921 S. 8th Ave. Stop 8116, Pocatello, ID 83209, USA
e-mail: altinich@isu.edu

Research in audiovisual speech perception has demonstrated that sensory factors such as auditory and visual acuity are associated with a listener's ability to extract and combine auditory and visual speech cues. This case study report examined audiovisual integration using a newly developed measure of *capacity* in a sample of hearing-impaired listeners. Capacity assessments are unique because they examine the contribution of reaction-time (RT) as well as accuracy to determine the extent to which a listener efficiently combines auditory and visual speech cues relative to independent race model predictions. Multisensory speech integration ability was examined in two experiments: an open-set sentence recognition and a closed set speeded-word recognition study that measured capacity. Most germane to our approach, capacity illustrated speed-accuracy tradeoffs that may be predicted by audiometric configuration. Results revealed that some listeners benefit from increased accuracy, but fail to benefit in terms of speed on audiovisual relative to unisensory trials. Conversely, other listeners may not benefit in the accuracy domain but instead show an audiovisual processing time benefit.

Keywords: audiovisual speech integration, hearing impairment, capacity, processing speed, speech reading, lip-reading

INTRODUCTION

While a listener's hearing ability certainly influences language performance, decades of research has revealed that cues obtained from the visual modality affect speech recognition capabilities (e.g., Sumby and Pollack, 1954; McGurk and MacDonald, 1976; Massaro, 2004). One common example is the classic *McGurk effect* in which incongruent or mismatched cues from the visual modality (e.g., auditory /ba/plus visually articulated "ga") influence auditory perception. Similarly, being able to see a talker's face under degraded listening conditions has been shown to facilitate both accuracy (Sumby and Pollack, 1954) and speed (e.g., Altieri and Townsend, 2011) compared to auditory-only recognition.

Auditory perceptual abilities are also associated with performance in the visual modality, as well as multisensory integration skills (Grant et al., 1998; Erber, 2003).

HEARING LOSS AND MULTISENSORY SPEECH CUES

High frequency hearing-loss

Research has consistently indicated that face-to-face communication capabilities are impacted less by hearing loss compared to auditory-only perception. The frequency range of hearing loss also influences audiovisual integration and social conversational ability. While high-frequency hearing loss at frequencies greater than 1000 Hz has an adverse effect on auditory-only perceptual abilities, audiovisual perceptual skills appear to be less adversely affected, as noted by Erber (2002, 2003) among others (e.g., Danhauer et al., 1985). This observation is noteworthy considering that a significant proportion of older adults experience a progressive high-frequency hearing loss commonly known as *presbycusis*. Prototypical audiograms indicative of presbycusis remain generally flat in the frequency range up until approximately 1000 Hz and slope progressively downward at higher frequencies. (As depicted in **Figure 1**, we shall refer to an audiogram showing evidence for only high-frequency hearing loss as a "sloping" audiogram). Importantly, people with high-frequency hearing loss have been reported to generally retain the ability to obtain low-frequency cues from the auditory signal including: manner of articulation, voicing, nasality, and vowel information (e.g., Erber, 2003). However, the perception of high-frequency speech sounds, such as fricatives (e.g., /∫/), is affected to varying degrees.

For adults with hearing loss, being able to see a talker's face can therefore prove exceedingly helpful in terms of enhancing accuracy. Facial movements—especially those associated with high-frequency sounds such as place of articulation—can "fill in" for auditory speech cues that have become degraded (Erber, 1975). As an example, distinguishing between a bilabial vs. an alveolar stop ("ba" vs. "da") is often straightforward in the visual modality, but often proves difficult auditorially. This becomes most noticeable when the quality of the auditory input is poor due to a noisy listening environment, hearing loss, or a combination of these factors.

Low-frequency hearing loss

Relationships between low-frequency hearing acuity and speech recognition have also been reported. Erber (2002, 2003) reported that "normal-hearing listeners"—those with low-frequency hearing thresholds better than 20 dB HL—have little difficulty hearing spoken words at a normal conversational distance and a volume level of 70 dB HL. Listeners with hearing loss greater than or

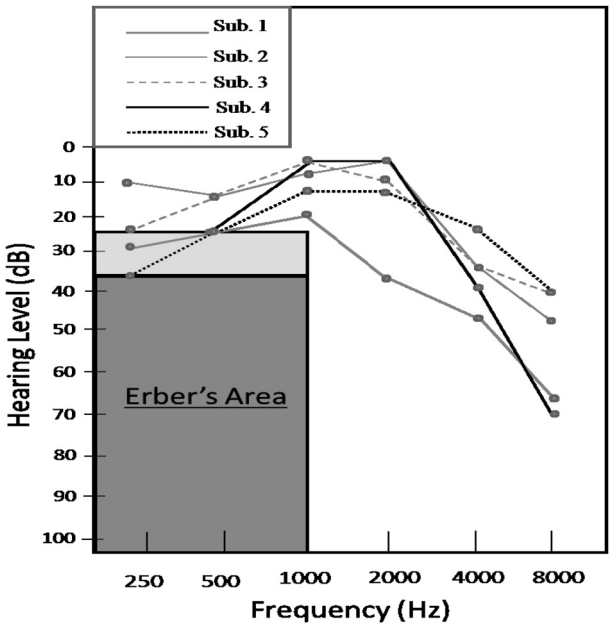


FIGURE 1 | Audiograms obtained from the five listeners with hearing loss. The gray area represents Erber's area, while the light gray area immediately above indicates a range of mild to moderate low frequency hearing impairment.

equal to approximately 25 dB HL often fail to recognize auditory linguistic cues about manner of articulation and vowels to varying degrees, while listeners with thresholds higher than 50 dB HL usually fail to accurately perceive any speech sounds through the auditory modality. Research has hence indicated that listeners with "flat audiograms," evidencing both low and high-frequency hearing loss may be poor integrators of audiovisual speech signals. This is ostensibly due to the fact that such listeners have a significantly reduced ability to isolate cues from the auditory signal related to voicing, nasality, vowel quality, and low frequency information. Listeners with low-frequency thresholds falling between 20 and 50 dB HL will predictably show significant variability in their ability to not only extract cues from the auditory speech signal, but combine them with complementary or redundant visemes. Taken together, the degree of hearing loss and pattern of errors in the unisensory modalities should contribute to a listener's ability to integrate multisensory cues information (e.g., Grant et al., 1998).

ASSESSING AUDIOVISUAL SPEECH INTEGRATION

We will implement the capacity assessment measure (Altieri et al., 2014b), known as *C_I(t)*, that compares both accuracy and reaction times (RTs) to parallel independent race model predictions derived from auditory and visual-only speech recognition trials. We therefore argue that a complete procedure for assessing integration ability should include multiple relevant dependent measures, namely accuracy and RTs. Comparisons between obtained data and predictions derived from statistical models that serve as a null-hypothesis will also be included (*Independent Race Models*; Miller, 1982). A RT-only measure of capacity, *C(t)*, (see Townsend and Nozawa, 1995) will also be assessed for each listener to separately diagnose RT capabilities. These methodologies are described further in Supplementary Material and by Altieri et al. (2014b). Generally, when capacity violates independent race model predictions at any time point (capacity $< ½$ or capacity > 1; Townsend and Wenger, 2004; Eidels et al., 2011; Otto and Mamassian, 2012) it indicates dependencies between auditory and visual modalities, and the presence of "integration." When independent predictions are not violated, it still may indicate that listeners benefit from visual information in a statistical manner simply due to the availability of a greater number of cues (i.e., AV vs. A or V-only). For instance, if a listener "misses" auditory phonemic cues, they still have the opportunity to obtain a relevant cue from the visual modality even if the visual modality does not facilitate auditory processing *per se*.

The predictive power of capacity is becoming increasingly established. First, research has shown that capacity is a superior predictor of cognitive performance, neural functioning, and recognition capabilities compared to mean accuracy and mean RTs (Wenger et al., 2010). Additionally, even RT-only capacity measures have been demonstrated to be predictive of multisensory gain in terms of accuracy (Altieri and Townsend, 2011), as well as EEG measures of audiovisual integration (Altieri and Wenger, 2013) and multisensory learning (Altieri et al., 2014a). Finally, research using a sample of listeners without reported hearing loss showed that *C(t)* and *C_I(t)* scores were associated with l pure tone thresholds as well as traditional accuracy-based assessments of audiovisual gain (Altieri and Hudock, 2014).

This study will illustrate how speed-accuracy tradeoffs occur by implementing a novel procedure for measuring a listener's integration skills that was recently applied to a group of normal-hearing listeners, and those with only mild hearing loss (Altieri and Hudock, 2014). This study here will go a step further by assessing integration efficiency (i.e., capacity) in hearing impaired listeners. The five case studies consist of listeners with self-reported hearing loss of different ages, and with varying degrees of high and low-frequency hearing-loss as measured by auditory pure-tone thresholds. In the following experiments, differences in audiovisual processing speed or accuracy compared to unisensory performance will be the primary means used to measure capacity qua integration efficiency. Our aim was to illustrate how listeners with high or low-frequency hearing loss can systematically differ from each other in their ability to benefit from combined audiovisual cues in the accuracy and processing time domains. In doing so, we sought to identify how tradeoffs in speed and accuracy occur in listeners with prototypical audiograms evidencing high or low-frequency hearing loss.

Hypotheses

Listeners with auditory sensory deficits should adopt certain predictable strategies when processing audiovisual speech stimuli to maximize multisensory benefit. In terms of speed-accuracy strategies, some individuals may be slower on audiovisual trials relative to independent race model predictions in order to take advantage of visual speech cues. However, they may also be substantially more accurate and potentially show evidence for super-capacity and efficient integration (i.e., in Experiment 2).

Additionally, we predict that these listeners should show greater audiovisual gain in open-set sentence recognition tasks (i.e., in Experiment 1). This scenario should often emerge in listeners with mild low-frequency and moderate high-frequency hearing difficulty. This type of hearing loss leads to mild degradation of certain vowel cues, and high-frequency information related to place of articulation. In these cases, auditory accuracy should be bolstered by visual speech cues. In fact, Altieri and Hudock (2014) showed that *C_I(t)* was correlated with low-frequency thresholds.

Second, we predict a larger RT-capacity gain (*C(t)*) for listeners with mild to moderate high-frequency hearing loss, but normal low-frequency thresholds. This is because auditory-only processing in these listeners should be slower but not significantly less accurate compared to normal-hearing listeners (due to the availability of more low-frequency vowel cues). This will, however, allow for complementary visual cues to facilitate speech recognition in the processing time domain especially. Generally, there should be a range in which auditory-only recognition becomes difficult, but enough cues are present in the signal to permit the combination of redundant and complementary visual cues to facilitate accuracy, perhaps in addition to speed. In future research, we predict that subjects with hearing loss will generally show superior integration skills compared to normal-hearing listeners, either in the speed, accuracy domain, or both.

METHODS

PARTICIPANTS

This study analyzed data obtained from five listeners recruited from the Idaho State University campus in Pocatello, ID that demonstrated hearing impairment on their audiogram which was obtained prior to the study (average pure tone threshold $\geq$25 dB SPL). The measured hearing loss could either occur for low-frequencies, high-frequencies, or both. Each of the five participants was a native speaker of an American English dialect. Each participant reported normal or corrected 20/20 vision[1]. The same listeners participated in both Experiments 1 and 2. This study was approved by the Idaho State University Human Subjects Committee, and each participant was paid 10 dollars per hour for their participation. This study required approximately 60 min to complete. Participant information, such as average low and high-frequency hearing thresholds, gender, and age is shown in **Table 1**.

AUDIOMETRIC TESTING

Pure-tone hearing thresholds were obtained for each volunteer prior to participation in this study using an *Ambco 1000 Audiometer*. Hearing thresholds were obtained in a sound attenuated chamber. Thresholds were obtained for 250, 500, 1000 (low frequencies), and, 2000, 4000, and 8000 Hz (high-frequencies; Erber, 2003) tones separately to each ear using headphones. For each frequency, thresholds were obtained via the presentation of a continuous tone. The following standard staircase procedure was used: when the listener identified the tone correctly by button press, the sound level was reduced 10 dB. If they failed to correctly indicate the presence of the tone, the decibel level was raised by 5 dB on the subsequent presentation.

Table 1 | Information for each of the five listeners, including average low and high-frequency pure tone threshold.

Participant	Age	Gender	Low frequency	High frequency	Hearing loss	Hearing aid
1	63	M	25	50	Sensory neural	Yes
2	22	F	8	35	Conductive	No
3	24	M	15	32	Sensory neural	No
4	60	M	17	37	Sensory neural	No
5	72	F	25	25	Sensory neural	No

EXPERIMENT 1: OPEN-SET SENTENCE RECOGNITION

Stimuli

The sentence stimuli used in Experiment 1 consisted of 75 sentences obtained from a database of recorded audio-visual sentences from the CUNY database (Boothroyd et al., 1985). Each of the sentences was spoken by a female talker. The stimulus set included 25 audiovisual, 25 auditory, and 25 visual-only sentences. The stimuli were obtained from a laser video disk and rendered into a 720 × 480 pixel video, digitized at a rate of 30 frames per second. Each stimulus was displayed on a standard Dell computer monitor with a refresh rate of 75 Hz. The auditory track was removed from each of the sentences using Adobe Audition for the visual-only sentences, and the visual component was removed for the auditory-only block. The sentences were subdivided into the following word lengths: 3, 5, 7, 9, and 11 words with five sentences for each length for each stimulus set (Altieri et al., 2011). This was done because sentence length naturally varies in conversational speech. Sentences were presented randomly for each participant, and we did not provide cues regarding to sentence length or semantic content. The sentence materials are displayed in Supplementary Material. To avoid ceiling performance, the auditory component of the signal was degraded using an 8-channel sinewave cochlear implant simulator (AngelSim: http://www.tigerspeech.com/). Consistent with Bent et al. (2009), we selected the following settings for the CI simulator: band pass filters were selected to divide the signal into eight channels between 200 and 7000 Hz (24 dB/octave slope), and a low pass filter was used to derive the amplitude envelope from each channel (400 Hz, 24 dB/Octave slope). Cochlear implant simulation with this number of channels generally leads to accuracy scores of approximately 70% words correct in young normal-hearing listeners in sentence recognition. Furthermore, it yields similar accuracy as multi-talker babble background noise (Bent et al., 2009).

Procedure

Accuracy data from the 75 audiovisual (25), auditory-only (25), and visual-only (25) sentences listen in Supplementary Material were obtained from each participant. Trials were presented in

[1]Altieri and Townsend (2011) showed that substantial degradation of the visual speech signal (e.g., darkening by 90% using Final Cut Pro) is necessary to measurably affect recognition capabilities. This is because the visual cues used to detect the speech signal, such as visual place of articulation, are highly salient. Therefore, in listeners with normal or corrected vision, visual acuity should not have been a factor influencing audiovisual integration skills.

separate blocks consisting of 25 audiovisual, 25 auditory, and 25 visual-only trials. The order of audiovisual, auditory, and visual-only block presentation was randomized across participants in an effort to avoid order effects. The stimuli in both experiments were presented to the participants using E-Prime 2.0 (http://www.pstnet.com/eprime.cfm) software.

Participants were seated in a chair approximately 24 inches from the monitor. Each trial began with the presentation of a black dot on a gray background, which cued the participant to press the space bar to begin the trial. Stimulus presentation began with the female talker speaking one of the sentences. After the talker finished speaking the sentence, a dialog box appeared in the center of the monitor instructing the participant to type in the words they thought the talker said by using a keyboard. Each sentence was given to the participant only once, and feedback was not provided on any of the trials. Scoring was carried out in a manner similar to the protocol described by Altieri et al. (2011). Whenever the participant correctly typed a word, then that word was scored "correct." The proportion of words correct was scored in each sentence. Word order was not a criterion for a word to be scored correctly, and typed responses were manually corrected for misspellings. (Upon inspection of the data, participants did not switch word order in their typed responses.). As an example, for the sentence "Is your sister in school," if the participant typed "Is the. . ." only the word "Is" would be scored as correct, making the total proportion correct equal to 1/5 or 0.20.

EXPERIMENT 2: SPEEDED WORD RECOGNITION: CAPACITY ANALYSIS

Stimuli

The stimulus materials consisted of audiovisual movie clips consisting of two female talkers. The stimuli were obtained from the Hoosier Multi-Talker Database (Sherffert et al., 1997). Two recordings of each of the following monosyllabic words were obtained from two female talkers: *Mouse*, *Job*, *Tile*, *Gain*, *Shop*, *Boat*, *Page*, and *Date*. These stimuli were drawn from similar studies carried out by Altieri and Townsend (2011), and also by Altieri and Wenger (2013). The auditory, visual, and audiovisual movies were edited using *Adobe After Effects*. Each of the auditory files was sampled at a rate of 48 kHz (16 bits). Each movie was digitized and rendered into a 720 × 480 pixel clip at a rate of 30 frames per second. Similar to Experiment 1, the auditory component signal was degraded using the 8-channel sinewave cochlear implant simulator. The duration of the auditory, visual, and audiovisual files ranged from 800 to 1000 ms. A previous report demonstrated that the variation in the duration of the movies did not influence RTs; rather, linguistic factors such as the confusability of the auditory and visual phonetic cues proved to be a major factor affecting processing speed (Altieri and Wenger, 2013). For example, the words "job" and "shop" were difficult to distinguish visually, and hence, visual-only RTs for these stimuli were slower and less accurate compared to the other words. Conversely, "boat" and "gain" were significantly easier to distinguish due to the difference in place of articulation.

Procedure

The audiovisual, auditory, and visual only trials were presented randomly in one block. There were a total of 128 audiovisual trials (64 spoken by each talker, where each of the 8 words was repeated 8 times per talker), 128 auditory-only trials, and 128 visual-only trials, for a total of 384 experimental trials. This portion of the experiment required 20–30 min to complete. Experimental trials began with a white dot on a gray background appearing in the center of the monitor. Each trial consisted of auditory-only, visual-only, or audiovisual stimuli. Auditory stimuli were played at a comfortable listening volume (approximately 70 dB SPL) over *Beyer Dynamic-100* Headphones.

Responses were collected via button press using a keyboard. Each of the buttons, 1–8, was arranged linearly on the keyboard and was labeled with a word from the stimulus set. Participants were instructed to press the button corresponding to the word that they judged the talker to have said as quickly and accurately as possible. Responses were timed from the onset of the stimulus on each trial. Inter-trial intervals randomly varied on a uniform distribution between 750 and 1000 ms. On auditory-only trials, participants were required to base their response solely on auditory information, and on visual-only trials participants were required to lip-read to make the speeded response. Auditory-only trials were played with a blank computer screen. Similarly, visual-only trials were played without any sound coming from the speakers. Each listener received 48 practice trials at the onset of each experimental block to assist with learning the response mappings on the keyboard.

RESULTS AND DISCUSSION

GENERAL SUMMARY: RT AND ACCURACY

The results from the open-set sentence recognition experiment (Experiment 1) are shown in **Table 2** and the results for the speeded-word recognition task (Experiment 2) are shown in **Table 3**. The tables show auditory- (A) and visual- (V) only percent correct for the CUNY sentences, respectively, as well as the predicted $\hat{p}(AV) = p(A) + p(V) - p(A) * p(V)$ and obtained audiovisual (AV) scores. The actual AV Gain scores are also displayed ($AV_{Gain} = p(AV) - max\{p(A), p(V)\}$; cf. Altieri and Wenger, 2013). For comparison purposes, the mean auditory, visual-only and audiovisual accuracy scores and the predicted and obtained AV integration scores are displayed for normal-hearing participants selected from another study (Altieri and Hudock, 2014).

Table 2 reveals that the two listeners with the lowest auditory-only accuracy (i.e., 1 and 2) showed the highest audiovisual gain, as predicted. Both listeners yielded gains that were greater than 2.5 SDs from the control participants. Interestingly, **Table 3** showed

Table 2 | CUNY sentence recognition scores for each listener.

Listener	A	V	Predicted AV	Obtained AV	AV gain
Average	78.40(7.20)	14.50(7.80)	81.40(6.80)	95.00(3.10)	16.60(6.80)
1	57	13	63	95	38
2	51	14	58	93	42
3	63	0	63	91	28
4	59	13	64	88	29
5	71	11	74	94	23

Table 3 | Speeded word recognition accuracy scores, mean RTs, and standard deviations (parentheses).

Listener	A	V	Predicted AV	AV	Max(*C_I(t)*)	Max(*C(t)*)	AV_RT	A_RT	V_RT
Average	98	73	99	98	1.28	1.40	1812 (284)	1823 (291)	2432 (514)
1*	55	30	69	95	5.40	0.49	1602 (244)	1993 (601)	1509 (353)
2	97	77	98	99	1.34	1.19	1790 (529)	1775 (549)	2124 (489)
3	99	80	99	100	0.84	0.51	2091 (521)	1875 (473)	3297 (1301)
4	100	70	100	99	1.41	2.47	2258 (899)	2332 (627)	3312 (1621)
5	98	51	99	99	1.10	1.20	2126 (545)	2129 (542)	2877 (811)

The standard deviation (SD) for the normal-hearing listeners was calculated across individual listeners. The "" indicates lower auditory and visual-only accuracy for this listener, along with considerably higher C_I(t).*

very low auditory-only accuracy for Participant 1. This listener reported difficulty in distinguishing several words that differed on key high-and low-frequency characteristics such as certain consonants (e.g., "job" vs. "shop"), and vowels, respectively (e.g., "mouse" vs. "boat"). Another feature of this listener was the comparatively fast visual-only responses. It appears that this listener responded "fast" on visual-only trials in order to get the trial finished quickly, perhaps because he was aware of the difficulty in accurately identifying content. However, when visual cues were combined with auditory, the listener was able to slow down in order to effectively merge the information with auditory cues. These findings will be further explored in the subsequent capacity analysis.

Next, the average capacity data for a group of normal-hearing participants are displayed in **Figure 2**. Together, these values denote the capacity-integration measures obtained at each time point. The RT-only *C(t)* values are displayed in the left panel, and the RT-accuracy *C_I(t)* values on the right. The dotted lines indicate one standard error (SE) of the mean.

Figure 3 shows capacity separately for hearing-impaired listeners 1 through 5. In each plot, the thick line shows *C_I(t)*, while the lighter line shows the RT-only *C(t)*. The actual capacity value at a time point is shown on the y-axis and the RT is displayed on the x-axis. To facilitate discrimination between strong integration ability (super-capacity) vs. poorer integration ability (limited capacity), two separate bounds can be found in each panel. First, limited capacity is defined by the dotted line at capacity = ½; the reason is that if processing resources were evenly divided between channels, then we would predict that the energy expended on AV trials would be half of the sum expected from A plus V efficiency (Townsend and Nozawa, 1995). Such a scenario would result if multisensory interactions between auditory and visual modalities were present, but the visual signal inhibited auditory recognition (e.g., Eidels et al., 2011). A useful heuristic bound separating unlimited and super-capacity corresponds to capacity = 1; the reason is that the race model inequality predicts capacity equal to 1 if processing on AV trials equals the sum of A and V processing (Townsend and Nozawa, 1995).

Crucially, the difference between *C_I(t)* and *C(t)* at any specific point, along with accuracy, provides information about a participant's processing strategy and integration ability (Altieri et al., 2014b). This is because *C(t)* furnishes information about speed (is the listener able to take advantage of visual speech information in terms of speed?). Also, obtained AV accuracy measured in comparison to the formula, $\hat{p}(AV) = p(A) + p(V) - p(A) * p(V)$ furnish information about whether integration is efficient or inefficient in the accuracy domain. Suppose *C_I(t)* is greater than 1 indicating efficient integration. Further, suppose that *C(t)* is also greater than 1 and accuracy equals independent race model predictions. This scenario would indicate that the listener is an efficient integrator due to the ability to take advantage of processing time on audiovisual trials rather than accuracy *per se*. Analogous logic applies to other scenarios in which the listener may show evidence for $C\_I(t) > 1$, but slows down to obtain higher than predicted audiovisual accuracy ($C(t) < 1$). We shall now discuss capacity results for the individual hearing-impaired listeners by examining *C_I(t)*, *C(t)*, and the discrepancy between predicted and obtained audiovisual accuracy shown in **Table 3**.

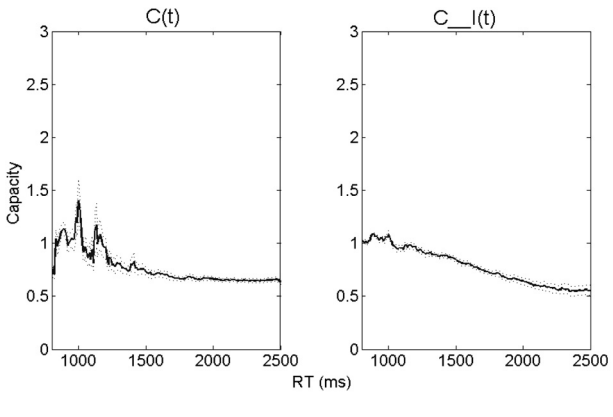


FIGURE 2 | A similar figure is displayed in Altieri and Hudock (2014), after some listeners with mild hearing impairment (thresholds > 25 dB) were removed (*n* = 39). It shows mean capacity-integration efficiency measures (thick solid line = Capacity/Integration Efficiency). The left panel shows *C(t)*, and the right panel shows *C_I(t)*. The dotted lines show one standard error (SE) of the mean.

Case 1

This first case involves a 63 year-old male with mild-moderate sudden-onset bilateral high and low-frequency hearing loss of unknown origin. The average pure tone threshold for the low frequency tones (250–1000 Hz) was approximately 20–25 dB HL, while the average threshold for the high frequency tones

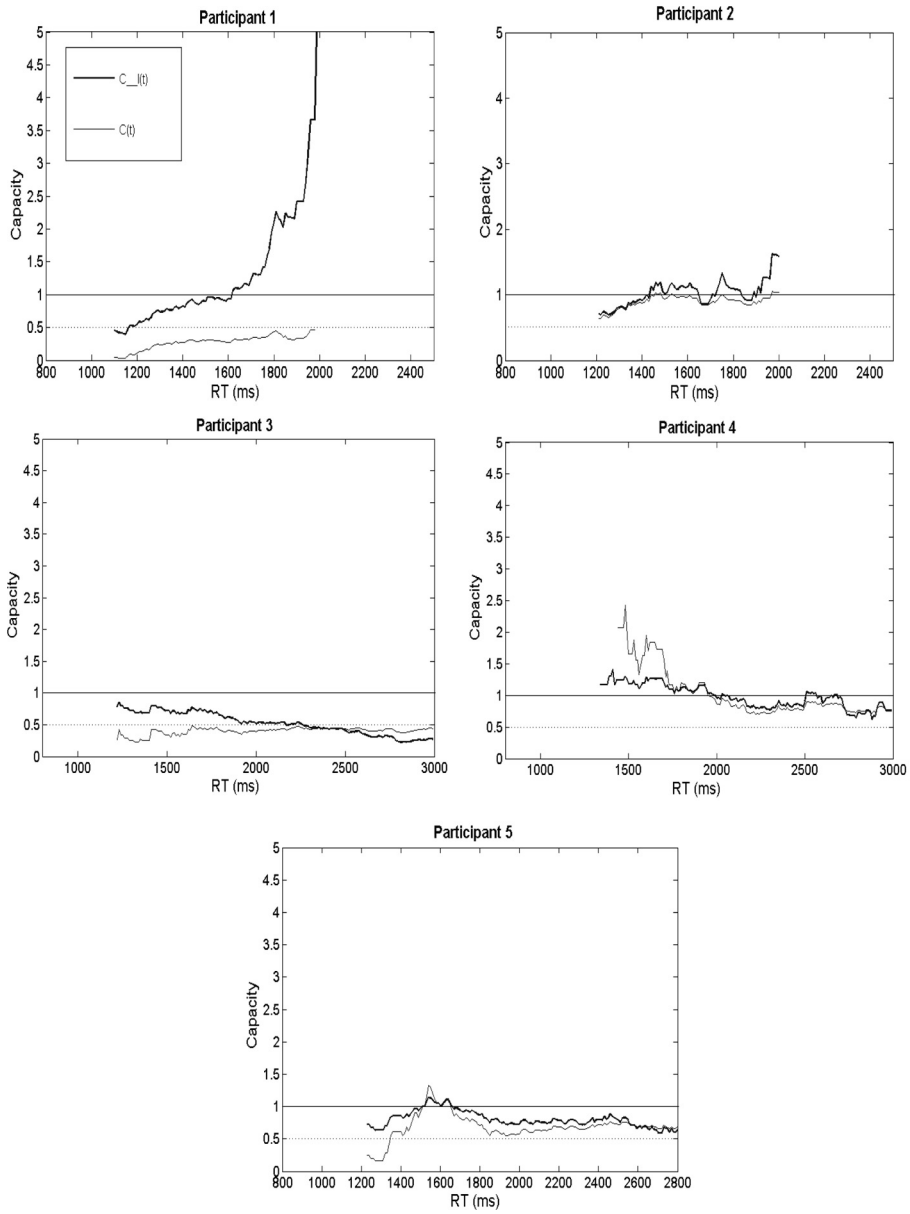


FIGURE 3 | Capacity (*C(t)* and *C_I(t)*) plotted individually for each of the five listeners.

(2000–8000 Hz) was 45–50 dB HL. This listener wears bilateral hearing aids to facilitate everyday perception, but did not use a hearing aid during testing.

The upper left hand panel of **Figure 3** (labeled "Participant 1") shows both capacity measures (accuracy was reported in **Table 3**). Remarkably, *C(t)* was extremely limited and less than ½ for the preponderance of time points. This finding is intriguing because it also shows that for each point in time, that Participant 1's *C(t)* was less than the average obtained from the group of normal-hearing volunteers (left panel of **Figure 2**). It indicates that Participant 1 slowed down on audiovisual trials compared to independent race model predictions. Integration measured in terms of speed (*C(t)*) was thus exceptionally poor for this listener. This change in speed appears to have resulted from a speed-accuracy tradeoff on visual-only trials since the listener was faster than average, but showed poor accuracy. However, when visual information was combined with the auditory modality, the listener proved capable of taking advantage of the combined auditory-visual information by slowing down to achieve better audiovisual accuracy. When both speed and accuracy were taken into account, Participant 1 displayed evidence for superior integration abilities across most times—particularly slower RTs. This listener elicited greater integration efficiency than normal-hearing participants, and for times greater than 1600 ms post stimulus onset, also demonstrated super-capacity. As observed from the data in **Table 3**, this higher than predicted audiovisual

accuracy (Predicted = 0.69; Obtained = 0.95) rather than speed drove integration ability.

While this listener had mild low-frequency hearing loss and thus generally poor performance in the auditory domain, he appeared to compensate by slowing down on audiovisual trials relative to visual-only trials in order to maximize accuracy—thus optimizing integration and residual perceptual matching strategies. Significantly, capacity was also reflected by the substantial audiovisual gain on sentence recognition. While auditory-only CUNY sentence recognition was poor due to hearing loss, gain scores were substantially high and only exceeded by Participant 2. This listener's integration skills predict strong face to face communication ability, resulting from the ability to obtain relevant cues from the auditory modality and combine them with visual information. Nonetheless, hearing aids likely facilitate speech communication by providing additional auditory cues, which should help speed-up information processing in difficult listening environments.

Case 2

Case number 2 involved a 22 year old female with bilateral high-frequency bone conductive hearing loss resulting from surgery as a toddler. This listener reported that she never used a hearing aid. The average low frequency threshold was 8 dB, while the average high-frequency threshold was approximately 35 dB HL.

The panel in **Figure 3** labeled "Participant 2" displays the capacity results. The RT only *C(t)* data show evidence for unlimited capacity across most time points, with the exception of very fast processing times. This indicates that capacity was consistent with independent race model predictions since *C(t)* was approximately unlimited. This finding indicated that Participant 2 was capable of statistically benefiting from visual cues in the RT domain. Furthermore, *C(t)* was higher compared to the average obtained from the normal-hearing listeners, and also higher than Participant 1's.

One explanation for this particular finding was that the degraded consonant information contributed to slower auditory-only responses and poor auditory-only sentence recognition capabilities; however, enough low-frequency vowel information was present for the listener to enable correct unisensory recognition in the context of a small set size, albeit at a slower pace. When both auditory and visual cues were present, this listener adequately matched the visual to the auditory cues to compensate for the deficits in auditory recognition, and hence, achieved borderline efficient integration. This processing speed difference was perhaps mirrored by the substantial accuracy gain achieved in open-set sentence recognition (**Table 2**). When combining the relative efficient integration in the time domain with accuracy levels in the forced choice task (Predicted = 0.98; Obtained = 0.99), overall integration skills indexed by *C_I(t)* shows evidence for unlimited capacity that was driven by RTs. Similar to *C(t)*, *C_I(t)* was greater than the average obtained from the normal-hearing listeners.

Case 3

Case number 3 was a 24 year old male with mild to moderate high-frequency sensory neural hearing loss—reportedly due to repeated noise exposure while serving in the military. This listener did not use a hearing aid. His audiogram showed that the average low frequency threshold in the better ear was 15 dB HL, while the average high-frequency threshold was approximately 30 dB HL.

Unlike the participant in case 2 who is of similar age and hearing ability, an overview of Participant 3's capacity measures revealed poor integration skills—in part due to poor visual-only skills as reflected in the CUNY sentence recognition task (see **Table 2**). Furthermore, this listener was slow but accurate on visual-only trials, and slower on audiovisual compared to auditory-only trials in the speeded word recognition task. These results perhaps indicate the presence of inhibition from the visual modality (Altieri and Townsend, 2011; Altieri and Wenger, 2013). Hence, for two listeners with similar demographics and thresholds, one observes differences in capacity that may be due to differences in hearing history, or other perceptual capabilities (e.g., visual-only). As one may observe in the panel labeled "Participant 3," both capacity measures are lower than the normal-hearing average, as well as Participant 2's for many time points. The *C(t)* hovers around the lower bound for fixed capacity indicating that this listener is "slow" when extracting visual place of articulation cues and filling in for degraded consonant information in the auditory modality. The results revealed similar auditory-only mean RTs for Participants 2 and 3, although significantly slower audiovisual RTs for Participant 3.

Crucially, the unitary *C_I(t)* measure was slightly higher than *C(t)*, although it was lower than 1, and lower than the bound on fixed capacity for his slowest RTs. The reason that the overall integration metric of *C_I(t)* showed evidence for slightly better integration compared to *C(t)* was that obtained accuracy levels approximated predicted scores in the speeded word task, making up for sluggish audiovisual RTs. Additionally, this participant showed substantial audiovisual gain on CUNY sentences, indicating the ability to benefit from visual information when combined with auditory speech cues under certain circumstances. Taken together, it appears that this participant has the ability to benefit from multisensory cues although he may fail to benefit in the processing time domain.

Case 4

Case number 4 included a 60 year old male with moderate age related high-frequency hearing loss. This listener was aware of his hearing loss, but did not use a hearing aid to facilitate language perception. His average low frequency threshold in the better ear (right) was 17 dB, while the average high-frequency threshold was recorded as 37 dB.

Similar to the listener from case 2, this listener revealed a strong correspondence between *C(t)* and *C_I(t)*. This participant showed evidence of unlimited capacity due to the fact that both capacity measures corresponded to independent model predictions for the vast majority of time points. Audiovisual responses appeared slightly faster than race model predictions for early recognition times. Audiovisual mean RTs were faster on average compared to auditory only RTs by approximately 100 ms, and also faster than visual-only RTs, which were quite sluggish. This listener's visual-only perception on both the CUNY sentence perception and speeded word recognition tasks were in the normal

range at 13 and 70% correct, respectively; however, the visual-only RTs suggest that he slowed down to achieve this accuracy level. Interestingly, the observation of super-capacity, and that the audiovisual trials were processed faster than auditory-only trials, indicates that visual information about place of articulation sped-up auditory recognition.

Similar to listener 2, this listener exhibited an integration profile that was consistent with race model predictions. Overall, his integration was superior to the normal-hearing average, which should often be true of listeners with mild to moderate hearing-loss. Despite the relatively poor auditory-only performance due to the loss of high-frequency cues, this listener's integration and lip-reading skills should facilitate face-to-face conversation enough to reduce or eliminate the need for a hearing aid.

Case 5

Case number 5 was a 72 year old woman with a flat audiogram showing evidence for bilateral mild hearing loss. Her average low and high-frequency hearing thresholds in the better ear were measured at approximately 20–25 dB HL. This listener reported being unaware of her hearing loss, and consequently, did not use a hearing aid.

Unlike Participant 1's, the capacity results showed slightly limed to unlimited capacity or mildly inefficient integration for responses slower than 1500 ms, but close to unlimited capacity for faster responses. This observation places this listener at odds with the group of normal-hearing listeners who showed super-capacity for faster processing times when comparing this participant's results with those shown in **Figure 2**. The mildly inefficient integration appears to suggest that this listener failed to benefit in terms of speed from visual cues (the fastest audiovisual responses did not differ from the fastest auditory-only RTs, and were only slightly faster than the visual-only RTs). Although audiovisual RTs were generally sluggish compared to independent model predictions, accuracy was at ceiling, and equal to predictions.

Overall, the measured *C_I(t)* was only marginally less than 1 for the preponderance of recognition times since accuracy made up for the moderate deficits in speed. In fact, since this listener's integration ability approximates the skills of normal-hearing listeners for most time points, she will likely be an effective face-to-face communicator without the use of a hearing aid, except in challenging conversational environments.

GENERAL DISCUSSION

The purpose of this study was to provide novel applications of a new capacity approach to identify loci of audiovisual speech integration abilities in hearing-impaired listeners. Specifically, this study extended recent capacity results using single point summaries (Altieri and Hudock, 2014) by illustrating how differences in auditory sensory acuity may be related to speech integration skills in listeners with different types of hearing impairment.

Of course, the picture appears somewhat complex as factors besides sensory acuity affect integration skills. Our results revealed how speech integration differs among hearing impaired individuals. Our results did demonstrate that older and younger listeners with different levels of hearing loss can yield similar capacity. For example, listeners 2 and 4 different in age (22 vs. 60 years old, respectively), and had different hearing loss etiology. In spite of these differences, integration skills as measured by accuracy and capacity were remarkably similar. On the other hand, listeners 1 and 5 were both over 60, and displayed similar audiograms; however, their integration skills differed both qualitatively and quantitatively. The upshot of these findings is that while relationships have been shown to emerge between sensory acuity and integration skills (e.g., Altieri and Hudock, 2014), pure tone thresholds are just one predictor of integration ability.

Cognitive factors also contribute to integration performance. In this report, however, we attempted to minimize the impact of higher cognitive factors at least, such as memory capabilities[2]. Yet, age may be one factor impacting integration abilities, as it has recently been shown to be correlated with capacity (Altieri and Hudock, 2014), and associated with poorer lip-reading (Sommers et al., 2005). One possibility is that many older listeners may effectively utilize the visual speech modality, but only in the context of sufficient auditory speech information. Therefore, we predict that older listeners with mild hearing loss may be efficient integrators—particularly in the accuracy domain. Audiovisual RTs, however, may be slower than independent model predictions in order to allow these listeners sufficient time to obtain cues and thus maximize accuracy. Interestingly, this speed-accuracy trade-off associated with aging has been consistently reported by Ratcliff and colleagues across a variety of cognitive and perceptual tasks (e.g., Ratcliff et al., 2004). The results provided by Participant 1 yields key evidence for this hypothesis. Nonetheless, because we utilized case study methodology, conclusive statements about the relationship between variables such as hearing loss etiology, age, and audiometric configuration and integration skills are difficult to ascertain. Critically, such integration strategies would be impossible to uncover using most current approaches for assessing integration which rely on accuracy-only (Braida, 1991; Grant et al., 1998; Massaro, 2004).

ASSESSING AUDIOVISUAL SPEECH INTEGRATION

Our findings show promise inasmuch as they indicate the importance of incorporating comprehensive speed-accuracy assessment measures. Processing speed is one critical variable predictive of information processing abilities, and one that is well-known to be adversely affected by aging and of course hearing ability. Unfortunately, speed has been overlooked as a viable measure of integration (see Altieri and Townsend, 2011; Winneke and Phillips, 2011). Before the capacity approach can be incorporated into future audiological assessment protocols, key developments seem to be in order. Obtaining normative data on integration skills using *C_I(t)* and *C(t)* will be necessary. Another important development will involve collecting larger data sets consisting

[2]All participants were employed or full time students and none of them reported any history of cognitive impairment or traumatic brain injury. Furthermore, each listener participated in a 5-min cued memory task involving a "study" and subsequent "test phase" (i.e., judging whether a picture was "old" or "new"). Each of the 5 participants scored 83% correct or better. The average score in a group of 84 participants (part of another study) was 91% correct with a standard deviation of 5.6%. Therefore, each listener was within 2.0 SDs of the mean of the larger sample of volunteers.

of hearing-impaired listeners with different audiometric configurations and different levels of cognitive functioning. Overall, capacity measures should prove important since audiologists do not comprehensively assess either visual or audiovisual processing capabilities in those suspected of hearing impairment, even though these skills are relevant for face-to-face communication capabilities.

SUMMARY AND CONCLUSION

This report provided a basis for a comprehensive methodological approach for examining speech integration in listeners with suspected hearing loss. Accuracy in open-set sentence recognition may be assessed subsequent to traditional audiometric testing. Next, the suggested approach would be to measure *C(t)* and *C_I(t)* using a closed-set speeded word recognition experiment to analyze the extent to which a listener benefits from multisensory cues relative to the predictions of independent models in which integration does not occur. Such a protocol will allow one to determine the locus of a listener's integration capabilities. For example, suppose a listener exhibits high *C_I(t)* in conjunction with low *C(t)*. This could indicate an inability to benefit from visual speech cues in terms of processing speed, but that slowing down may help one take advantage of visemes to achieve high accuracy.

Notwithstanding our findings, one may observe that while audiograms and auditory-only hearing ability may be associated with integration (Erber, 2002, 2003), individual differences in integration ability exist. Besides auditory sensory capabilities, other sensory or cognitive factors can influence integration ability. This includes age (Bergeson and Pisoni, 2004; Ratcliff et al., 2004; Sommers et al., 2005; Winneke and Phillips, 2011), complex visual-only perceptual abilities (i.e., face recognition), processing speed and related strategies, and working memory skills (Lunner et al., 2009). Interaction among these factors and how they relate to speech integration skills have only recently begun to be explored in a model-based way and therefore require considerable future investigation.

ACKNOWLEDGMENTS

The project described was supported by the INBRE Program, NIH Grant Nos. P20 RR016454 (National Center for Research Resources), and P20 GM103408 (National Institute of General Medical Sciences).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00678/abstract

REFERENCES

Altieri, N., and Hudock, D. (2014). Variability in audiovisual speech integration skills assessed by combined capacity and accuracy measures. *Int. J. Audiol.* doi: 10.3109/14992027.2014.909053. [Epub ahead of print].

Altieri, N., Pisoni, D. B., and Townsend, J. T. (2011). Some normative data on lip-reading skills. *J. Acoust. Soc. Am.* 130, 1–4. doi: 10.1121/1.3593376

Altieri, N., Stevenson, R. A., Wallace, M. T., and Wenger, M. J. (2014a). Learning to associate auditory and visual stimuli: capacity and neural measures of efficiency. *Brain Topogr.* doi: 10.1007/s10548-013-0333-7. [Epub ahead of print].

Altieri, N., and Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Front. Psychol.* 2:238. doi: 10.3389/fpsyg.2011.00238

Altieri, N., Townsend, J. T., and Wenger, M. J. (2014b). A measure for assessing the effects of audiovisual speech integration. *Behav. Res. Methods.* 46, 406–415. doi: 10.3758/s13428-013-0372-8

Altieri, N., and Wenger, M. (2013). Neural dynamics of audiovisual integration efficiency under variable listening conditions: an individual participant analysis. *Front. Psychol.* 4:615. doi: 10.3389/fpsyg.2013.00615

Bent, T., Buchwald, A., and Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *J. Acoust. Soc. Am.* 126, 2660–2669. doi: 10.1121/1.3212930

Bergeson, T. R., and Pisoni, D. B. (2004). "Audiovisual speech perception in deaf adults and children following cochlear implantation," in *The Handbook of Multisensory Processes*, eds G. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: MIT Press), 749–771.

Boothroyd, A., Hanin, L., and Hnath, T. (1985). *A Sentence Test of Speech Perception: Reliability, Set equivalence, and Short Term Learning (Internal Report RCI 10)*. New York, NY: City University of New York.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Q. J. Exp. Psychol.* 43A, 647–677. doi: 10.1080/14640749108400991

Danhauer, J. L., Garnett, C. M., and Edgerton, B. J. (1985). Older persons' performance on auditory, visual, and auditory-visual presentations of the Edgerton and Danhauer Nonsense Syllable Test. *Ear Hear.* 6, 191–197. doi: 10.1097/00003446-198507000-00004

Eidels, A., Houpt, J., Altieri, N., Pei, L., and Townsend, J. T. (2011). Nice guys finish fast and bad guys finish last: a theory of interactive parallel processing. *J. Math. Psychol.* 55, 176–190. doi: 10.1016/j.jmp.2010.11.003

Erber, N. P. (1975). Auditory-visual perception of speech. *J. Speech Hear. Disord.* 40, 481–492.

Erber, N. P. (2002). *Hearing, Vision, Communication, and Older People*. Clifton Hill, VIC: Clavis Publishing.

Erber, N. P. (2003). Use of hearing aids by older people: influence of non-auditory factors (vision, manual dexterity). *Int. J. Audiol.* 42, 2S21–2S25. doi: 10.3109/14992020309074640

Grant, K. W., Walden, B. E., and Seitz, P. F. (1998). Auditory-visual speech recognition by hearing impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 103, 2677–2690. doi: 10.1121/1.422788

Lunner, T., Rudner, M., and Ronnberg, J. (2009). Cognition and hearing aids. *Scand. J. Psychol.* 50, 395–403. doi: 10.1111/j.1467-9450.2009.00742.x

Massaro, D. W. (2004). "From multisensory integration to talking heads and language learning," in *The Handbook of Multisensory Processes*, eds G. A. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: The MIT Press), 153–176.

McGurk, H., and MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Miller, J. (1982). Divided attention: evidence for coactivation with redundant signals. *Cogn. Psychol.* 14, 247–279. doi: 10.1016/0010-0285(82)90010-X

Otto, T., and Mamassian, P. (2012). Noise and correlations in parallel decision making. *Curr. Biol.* 22, 1391–1396. doi: 10.1016/j.cub.2012.05.031

Ratcliff, R., Thapar, A., and McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *J. Mem. Lang.* 50, 408–424. doi: 10.1016/j.jml.2003.11.002

Sherffert, S., Lachs, L., and Hernandez, L. R. (1997). "The hoosier audiovisual multi-talker database," in *Research on Spoken Language Processing Progress Report No. 21*. Bloomington, IN: Speech Research Laboratory, Psychology, Indiana University.

Sommers, M., Tye-Murray, N., and Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear.* 26, 263–275. doi: 10.1097/00003446-200506000-00003

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 12–15.

Townsend, J. T., and Nozawa, G. (1995). Spatio-temporal properties of elementary perception: an investigation of parallel, serial and coactive theories. *J. Math. Psychol.* 39, 321–360. doi: 10.1006/jmps.1995.1033

Townsend, J. T., and Wenger, M. J. (2004). A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality serie. *Psychol. Rev.* 111, 1003–1035. doi: 10.1037/0033-295X.111.4.1003

Tye-Murray, N., Sommers, M., and Spehar, B. (2007). Audiovisual integration and lip-reading abilities of older adults with normal and impaired hearing. *Ear Hear.* 28, 656–668. doi: 10.1097/AUD.0b013e31812f7185

Wenger, M. J., Negash, S., Petersen, R. C., and Petersen, L. (2010). Modeling and estimating recall processing capacity: sensitivity and diagnostic utility in application to mild cognitive impairment. *J. Math. Psychol.* 54, 73–89. doi: 10.1016/j.jmp.2009.04.012

Winneke, A. H., and Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychol. Aging* 26, 427–438. doi: 10.1037/a0021683

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 November 2013; accepted: 11 June 2014; published online: 01 July 2014.
*Citation: Altieri N and Hudock D (2014) Hearing impairment and audiovisual speech integration ability: a case study report. Front. Psychol. **5**:678. doi: 10.3389/fpsyg.2014.00678*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 April 2015
doi: 10.3389/fpsyg.2015.00435

Audiovisual integration of speech in a patient with Broca's Aphasia

Tobias S. Andersen[1]* *and Randi Starrfelt*[2]

[1] Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark, [2] Department of Psychology, Center for Visual Cognition, University of Copenhagen, Copenhagen, Denmark

Lesions to Broca's area cause aphasia characterized by a severe impairment of the ability to speak, with comparatively intact speech perception. However, some studies have found effects on speech perception under adverse listening conditions, indicating that Broca's area is also involved in speech perception. While these studies have focused on auditory speech perception other studies have shown that Broca's area is activated by visual speech perception. Furthermore, one preliminary report found that a patient with Broca's aphasia did not experience the McGurk illusion suggesting that an intact Broca's area is necessary for audiovisual integration of speech. Here we describe a patient with Broca's aphasia who experienced the McGurk illusion. This indicates that an intact Broca's area is not necessary for audiovisual integration of speech. The McGurk illusions this patient experienced were atypical, which could be due to Broca's area having a more subtle role in audiovisual integration of speech. The McGurk illusions of a control subject with Wernicke's aphasia were, however, also atypical. This indicates that the atypical McGurk illusions were due to deficits in speech processing that are not specific to Broca's aphasia.

Keywords: audiovisual, speech perception, aphasia, Broca's area, multisensory integration

OPEN ACCESS

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Nicholas Altieri, Idaho State University, USA
Dawn Marie Behne, Norwegian University of Science and Technology, Norway

***Correspondence:**
Tobias S. Andersen, Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 321, 2800 Lyngby, Denmark
toban@dtu.dk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *26 November 2014*
Accepted: *27 March 2015*
Published: *28 April 2015*

Citation:
Andersen TS and Starrfelt R (2015) Audiovisual integration of speech in a patient with Broca's Aphasia. Front. Psychol. 6:435. doi: 10.3389/fpsyg.2015.00435

Introduction

Broca's area has long been known to be necessary for speech production as evidenced by the severe impairment to speech production, known as expressive, non-fluent or Broca's aphasia, caused by lesions to this area located in the ventrolateral prefrontal cortex (Broca, 1861). It has long been known that expressive aphasia can be dissociated from impairment of speech perception known as receptive, fluent or Wernicke's aphasia (Wernicke, 1874), caused by damage to Wernicke's area in the posterior superior temporal sulcus (STS). This dissociation may, however, not be entirely complete as speech perception under adverse conditions can be impeded in Broca's aphasics (Blumstein et al., 1977; Moineau et al., 2005). Both Broca's and Wernickes' area are located in the left hemisphere in about 90% of right-handers and 70% of left-handers (Knecht et al., 2000).

A role for Broca's area is also suggested by studies showing that this area is active not only during speech production, but also during speech perception, particularly when it involves lip-reading (Watkins et al., 2003; Ojanen et al., 2005; Skipper et al., 2005). Furthermore, speech perception also activates motor cortex, and this activation is articulator-specific so that labial articulations activate the lip region and lingual articulations activate the tongue region of the motor cortex (Pulvermüller et al., 2006). This effect is causal as shown by studies in which de-activation of premotor (Meister et al., 2007) and primary motor (Möttönen and Watkins, 2009) cortex using TMS impeded speech perception. As changes in the excitability of motor cortex due to speech perception

are correlated with activity in Broca's area (Watkins and Paus, 2004) this activation is likely to go via Broca's area (Möttönen and Watkins, 2012).

These findings suggest a role for Broca's area in speech perception partly echoing the claims made by the Motor Theory (MT) of speech perception (Liberman and Mattingly, 1985; Galantucci et al., 2006). According to the MT, in its strongest form (Liberman and Mattingly, 1985), the internal representation of phonemes is not based on auditory templates but instead on the motor commands issued to articulate the phonemes. Hence, according the MT, speech, perceived by hearing or lip-reading (Liberman, 1982), must be mapped onto motor commands, located in the motor system, to be understood.

The MT was fueled by the discovery of mirror neurons in pre-motor area F5 in non-human primates (Gallese et al., 1996). These neurons respond both when performing a goal-directed action and when seeing that action performed by others. As area F5 has been suggested to be the non-human primate homolog of Broca's area in humans (Rizzolatti and Arbib, 1998; Binkofski and Buccino, 2004) this discovery gave rise to the mirror neuron motor theory of speech perception: Mirror neurons in Broca's area may decode auditory and visual speech to articulatory actions and activate their representation in motor areas (Rizzolatti and Craighero, 2004). Several observations support this hypothesis. First, some mirror neurons can respond to hearing as well as seeing an action (Kohler et al., 2002), which seems like a necessary ability for neurons involved in speech perception. Second, mirror-neurons can respond not only to hand movements but also to communicative mouth movements (Ferrari et al., 2003).

Broca's area thus has a central role in audio-visual-motor integration according to the mirror neuron MT of speech perception. The integration of auditory and visual information in speech perception is evidenced by the congruency effect, which is a general facilitation when watching the interlocutor's face (Sumby and Pollack, 1954), and by the McGurk illusion (McGurk and MacDonald, 1976) in which the perception of a clearly perceived phoneme (e.g., /ba/) is altered (e.g., to /da/) when it is perceived dubbed onto an a face articulating an incongruent phoneme (e.g., /ga/). For this integration to occur the auditory and visual information must be mapped onto a common representation, which, according to the MT, is articulatory (Liberman, 1982). On the basis of the mirror neuron MT, Ramachandran and co-workers suggested, in a preliminary report, that Broca's area is necessary for audiovisual integration of speech (Ramachandran et al., 1999). In support of this hypothesis they reported results from a single Broca's aphasic that did not experience the McGurk illusion. This result has, to our knowledge, never been tested more fully, which is what we aim to do in the current study.

The strongest claim of the original MT is that the motor system is essential for speech perception, but even though fairly recent studies have argued for this (Fadiga et al., 2002; Meister et al., 2007), substantial evidence against it has also been reported (Lotto et al., 2009). The resolution may lie in a more complex and nuanced account of the MT in which the motor system is not necessary for speech perception but rather has a supplementary role (Scott et al., 2009).

This complexity is captured in dual-stream anatomical models in which the dorsal and ventral pathways link areas involved in speech perception with areas involved in speech production (Skipper et al., 2005, 2007; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Jääskeläinen, 2010). The dorsal stream projects from the posterior part of the STS, containing Wernicke's area, to premotor cortex, which projects further onto Broca's area. According to the classical Wernicke-Geschwind model this stream has a crucial role for feeding perceptual information to speech production centers although more recent evidence suggests it has a more complex role (Bernal and Ardila, 2009). Dual stream models of speech perception have suggested that the dorsal pathway also underlies perceptual processing of audiovisual speech based on articulatory representations (Skipper et al., 2005, 2007; Jääskeläinen, 2010). This suggests an important role for articulatory representations in audiovisual integration of speech, as the posterior STS is the area in the perceptual system that is most consistently found to be activated super-additively by audiovisual speech. The ventral stream projects from the anterior part of the STS to the ventrolateral prefrontal cortex, containing Broca's area (Bernstein and Liebenthal, 2014). Thus, although one dual-stream model hypothesizes that only processing in the dorsal pathway is based on articulatory representations (Skipper et al., 2005, 2007; Jääskeläinen, 2010) the ventral pathway may also be important for sensory-motor integration of speech. The complexity of sensory-motor interactions in this dual-stream network means that dysfunction of parts of the network may lead to subtle and complex effects in agreement with moderate versions of the MT. Yet, as both streams may ultimately project to Broca's area, the dual-stream model does not preclude that this area can have a necessary role in audiovisual integration of speech as suggested by Ramachandran et al. and the strong version of the mirror neuron MT.

The purpose of the current study is to test a patient suffering from Broca's aphasia in a standard auditory, visual, and audiovisual speech perception task. The hypothesis is that the patient should show no sign of audiovisual integration, as measured by the McGurk illusion and the congruency effect, if Broca's area is necessary for audiovisual integration of speech. Alternatively, audiovisual integration could be weak rather than completely eliminated, which would still support a supplementary role for Broca's area in audiovisual integration. As the McGurk illusion is subject to great individual variability, this alternative hypothesis is more difficult to test. Furthermore, as the strength of the McGurk illusion depends not only on the observer but also on the experimental setup (Magnotti and Beauchamp, 2014), especially the stimulus material, we also tested a patient with receptive aphasia and two healthy control participants to confirm that our experimental setup could induce the McGurk illusion.

Methods

Case Reports

Patient ML: Broca's Aphasia

ML is a right-handed, male native Danish speaker, who, at the time of testing was 47 years old. ML suffered an ischemic

stroke in 2006, at the age of 42, for which he received trombolysis treatment. The stroke affected the anterior 2/3 of the territory of the Middle Cerebral Artery (MCA) in the left hemisphere, including both cortex and white matter. A computerized tomography (CT) scan (see **Figure 1**) a week post-stroke, showed that both the inferior and middle frontal gyri were affected, and that there was a smaller cortical affection laterally in the superior frontal gyrus. There was also loss of substance frontotemporally around the Syvian Fissure. The entire insular cortex was affected, and the infarction stretched rostrally in centrum semiovale and medially in the left basal gangila. There was Wallerian degeneration with atrophy of the left cerebral penduncle and left mesencephalon, and a compensatory widening of the left lateral ventricle. The entire inferior frontal gyrus, which contains Broca's area, was thus affected by the lesion.

Initially following his stroke, ML had global aphasia, right hemianopia, and right-sided hemiparesis affecting the arm and leg. Neuropsychological examination 4 months post injury showed that ML was able to say "yes" and "no," as well as his own name and the name of his two children, but nothing else. At this time, he had severe apraxia of speech, and his receptive language abilities were also affected, while his understanding of gestures was comparatively good. His reading and writing was also severely affected. In 2007, ML was part of a 4 months intensive rehabilitation programme at the Center for Rehabilitation of Brain Injury (CRBI) in Copenhagen, and he subsequently continued cognitive and speech training for some hours per week at CRBI until the end of 2009. He was described as a resourceful and highly motivated person, who has benefitted greatly from the rehabilitation efforts. At the last neuropsychological assessment performed at the CRBI in December 2009, ML was described as having normal attention and concentration, normal problem-solving abilities and normal visual processing/perception. His psychomotor tempo was difficult to evaluate because of his remaining hemiparesis. His learning and memory scores were somewhat below normal. For arithmetic, he was able to perform addition and subtraction on a normal level, but had difficulties with multiplication and division. His naming abilities had improved significantly, and he was able to name 28/30 nouns and 23/30 verbs correctly (on initial testing in 2006, he could name 1/30 and 1/30, respectively). At the time of the present investigation, ML's right arm was still paretic, while the paresis of the right leg had remitted to the degree that he can walk. His hemianopia was also remitted. ML has returned to his previous job as an auto mechanic, but now works reduced hours and performs simpler tasks than before (partly because of the arm paresis). The experimental investigation reported here was conducted in June 2011.

Patient MK: Wernicke's Aphasia

MK is a left-handed, male native Danish speaker, who at the time of testing was 42 years old. MK suffered a haemorrhagic stroke in the territory of the left MCA in 2007. A CT scan performed one month post injury (see **Figure 1**) showed an intraparenchymal hemorrhage in the resorbtion phase centered on the basal ganglia, and stretching fronto-tempo-parietally in the left hemisphere. There was a central component of fresh blood measuring about $60 \times 20 \times 40$ mm, with an oedema of about 7–10 mm surrounding

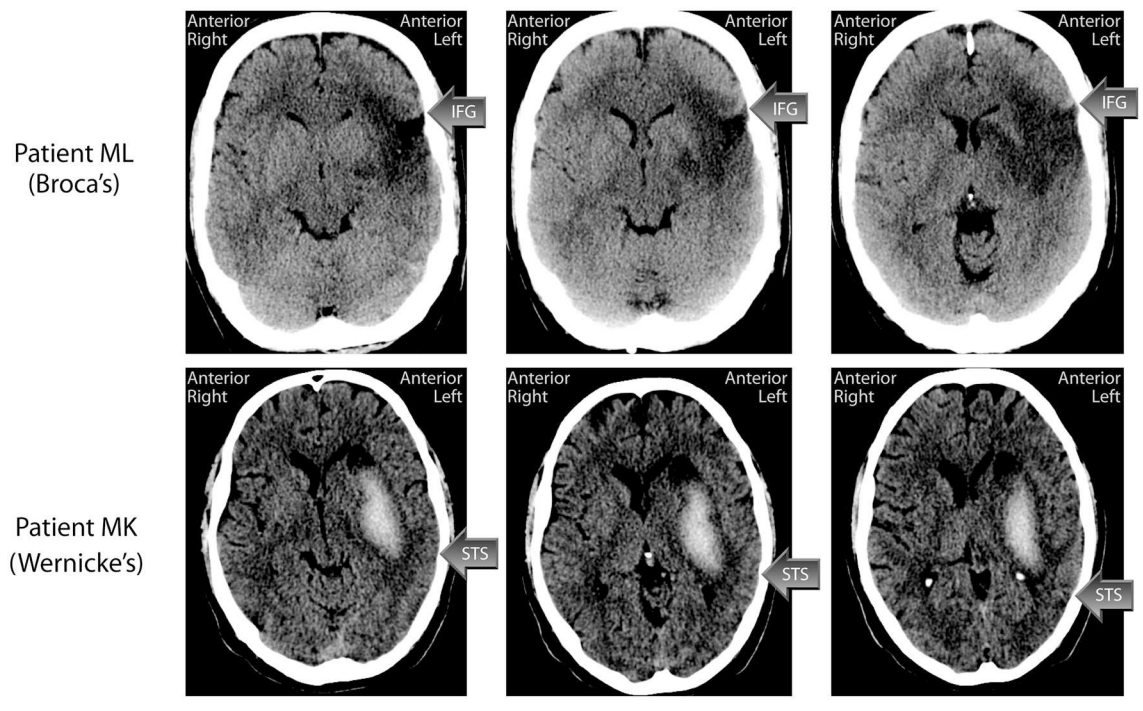


FIGURE 1 | CT-scans of the lesions of patients ML (Broca's aphasic) and MK (Wernicke's aphasic). Arrows indicate the approximate location of the inferior frontal gyrus (IFG) in Patient ML and the superior temporal sulcus (STS) in Patient MK. See main text for a description of the patients' lesions.

it. The hemorrhage stretched from the left mid temporal lobe to left centrum semovale and involved the left basal ganglia, although the caudate nucleus was spared. Cortex was also spared. The oedema affected the entire left temporal lobe, which makes the specific gyri and sulci difficult to discern. The left lateral ventricle was compressed, and the midline displaced about 3 mm toward the right. Hence the posterior part of the STS, important for audiovisual integration, may have been spared although the underlying white matter is likely to have been affected.

Initially, ML was unconscious, and spent the first 5 days in a respirator. During the first 40 days his Barthel Score increased from 0 to 49. After regaining consciousness, MK was globally aphasic and in the beginning his only response was opening his eyes when spoken to. His right arm and leg were paralytic. Soon after, he could respond to simple yes/no questions relatively correctly, and could point to some simple objects and colors. He had paralysis of the right side and a right-sided hemianopia. A logopaedic assessment two weeks post injury showed that language comprehension and production, as well as repetition were severely affected. Neuropsychological assessment, one month post injury, reported global aphasia, and problems in visuo-spatial tests. MK has been in intensive rehabilitation, first as an in-patient at Center for Neurorehabilitation—Filadelfia for six months, and later in the outpatient intensive programme at CRBI in Copenhagen for four months. A neuropsychological evaluation at CRBI in 2010 concluded that MK's visual working memory, visuo-spatial memory, and semantic processing (Pyramid and Palm trees) were relatively unaffected, while some difficulties were evident in more complex problem-solving tasks. He still suffered from moderate to severe aphasia. His confrontation naming was impaired, although greatly improved. His verbal comprehension was impaired, but he was noted to be relatively good at comprehending simple sentences. His repetition was also still impaired, as was his reading and writing. It was noted that MK often tried to write words he could not say and quite often succeeded in doing so. At the time of this investigation MK was receiving speech therapy at a community center. His left arm was still paretic, while the paralysis of the leg was remitted so that he could walk independently. MK was working part time as a practical aid at a school. The experimental investigation reported here was conducted in January 2014.

Aphasia Assessment

As part of the current investigation, ML and MK were assessed with the Danish version of the Western Aphasia Battery (WAB) (Kertesz, 1982) parts I–IV, to characterize their language abilities. Both patients provided written, informed consent according to the Helsinki declaration to participate in this study. The test results are presented in **Table 1**. ML's language deficit was classified as Broca's aphasia, and his Aphasia quotient was 72. MK's language deficit was classified as Wernicke's aphasia, and his Aphasia quotient was 64. The most important difference between the two patients was in the "Spontaneous speech" subsection of WAB, where, although their overall scores were almost identical, the two point difference in "Fluency" landed them on different sides of the diagnostic border between Broca's and Wernicke's aphasia according to the WAB diagnostic scoring system.

TABLE 1 | Subscores—Western Aphasia Battery.

	ML (Broca)	MK (Wernicke)
SPONTANEOUS SPEECH		
Functional content	9/10	8/10
Fluency	4/10	6/10
Total	*13/20*	*14/20*
COMPREHENSION		
Yes/no questions	57/60	48/60
Auditory word recognition	59/60	51/60
Sequential commands	28/80	35/80
Total	*144/200*	*134/200*
REPETITION		
Total	*56/100*	*48/100*
NAMING		
Object naming	57/60	45/60
Word fluency	11/20	6/20
Sentence completion	6/10	8/10
Responsive speech	9/10	5/10
Total	*83/100*	*64/100*
Aphasia quotient	***72***	***64***

Patient ML: Broca's Aphasia

At the time of the current investigation, ML's spontaneous speech was "telegraphic." He was, however, well able to make himself understood using a combination of words and gestures. He was able to mobilize nouns and a few verbs, but did not speak in full sentences with the exception of a few common phrases like "I don't know" or "It's difficult." As an example, in the picture description task from the WAB, ML mobilized 13 correct nouns/object names from the scene. Asked to try to describe the picture with full sentences, he added the definite article to the nouns (*a house, a man* etc.), but provided no verbs or function words. ML understood most questions and instructions either at first try or with a single repetition, and this is reflected in his relatively high score in the "Yes/no questions" and "Auditory word recognition" tests. In the more difficult "Sequential commands" test, he failed with more difficult instructions like "Use the book to point at the comb," but interestingly managed the more "ecologically understandable" instruction "Put the pen on the book and then give it to me." ML could repeat single words and simple sentences, but failed with sentences longer than four words. He named objects quite well, although his response-times are elevated.

In sum, ML's aphasia had remitted from Global aphasia to Broca's aphasia. He understood speech reasonably well, and his speech output was understandable but consisted mainly of nouns and very few verbs and function words.

Patient MK: Wernicke's Aphasia

MK's spontaneous speech was fluent, and he was able to construct some full, albeit simple, sentences like "Actually I feel very good." In the WAB picture description task, he referred to many of the objects in the picture using sentences like "There is a tree. There is also a flagpole." Sometimes, when MK could not find the right

word in speech, he could write it down correctly and then read it. He also sometimes resorted to English words when he could not find the Danish ones. In auditory verbal comprehension, he answered most questions correct (scoring 48/60), but failed relational (and grammatically challenging) questions like "Is a dog larger than a horse?" and "Does July come before May?" He could point to most images and objects in the auditory word recognition task, but had problems with left-right discrimination of body parts, and with pointing to the correct fingers. He was able to follow simple commands, but failed with longer sentences, particularly if the sequence of actions to be performed did not correspond to the word sequence. He always used the objects in the order mentioned even when this was incorrect (e.g., He responded correctly to "Point with the pen to the book" but not to "Point to the book with the pen"). He could repeat sentences up to four words correctly. In object naming he made quite a few errors (45/60 correct). Many of the responses were either phonological or semantic paraphasias ("spoon" for "fork"; "brushtooth" for "toothbrush"). A few of the words he could not name orally were written correctly (e.g., toothbrush). His responsive naming was at about the same level as his naming of objects.

In sum, MK's aphasia had remitted from Global aphasia to Wernicke's aphasia. He understood speech reasonably well, and his speech output was understandable although he had word-finding difficulties and some paraphasias.

Control Participants

In order to verify that auditory and visual stimuli were identified correctly and that the audiovisual stimuli could induce the McGurk illusion in healthy subjects we also tested two male native Danish speakers (JL, age 49, and PB, age 52) with no history of neurological disorders and self-reported normal hearing. The control participants were chosen to match the gender, approximate age, and native language of the aphasic patients.

Stimuli and Procedure

The stimuli were based on video (with audio) recordings of a male talker (one of the authors, TSA) pronouncing the bi-syllabic non-words /aga/, /aba/, and /ada/. The recordings were made using a Panasonic™ AG-HMC41E video camera, which recorded video at 25 frames per second with a resolution of 1920 × 720 pixels and audio at 44.1 kHz via an external DPA™ d:fine 4066-F microphone headset, which was not visible in the video recordings. The video was converted to QuickTime™ files with a resolution of 640 × 360 pixel to ensure efficient playback. Stimuli were presented using Matlab™ and the Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) on a MacBook™ Pro laptop computer and a pair of Genelec™ 6010A active speakers.

In each experiment, the subject was seated with his ears approximately 57 cm from the display. This was done by holding a length measure so that it extended from the plane of the speakers toward the subject and instructing the subject to adjust his position so that his ears were 57 cm from the speakers. The participants were observed to ensure that they did not move their head substantially but their position was not restrained during the experiment. The sound level was approximately 56 dB(A) measured using a sound level meter at the approximate location of the participant's head. This rather low sound level was chosen because the McGurk illusion is stronger when the sound intensity is low (Sekiyama and Tohkura, 1991; Andersen et al., 2001).

The stimuli were auditory, visual, or audiovisual. Auditory stimuli were presented with a still photograph of the talker's face visible. Visual stimuli were presented with no sound and the task was to lip-read. Audiovisual stimuli consisted of video and sound recordings presented synchronously. The sound and video could either be one of the three natural congruent combinations or any of the six possible incongruent combinations (cf. **Figure 2**) of the auditory and visual stimuli. The incongruent combinations were created in Final Cut Pro by dubbing the incongruent acoustic speech recording onto the video aligning the consonant burst of the incongruent acoustic speech recording with the original congruent acoustic speech recording.

After a stimulus had played the response options "B," "BG," "G," "BD," and "D" were shown in large letters on the screen. The patients could respond by pointing to one of the response categories displayed on the screen, repeating what he had heard or declining to give an answer. We allowed for these different ways of responding in order to accommodate the patients who could have problems pointing to the response categories due to problems with reading, problems with repeating what they heard due to apraxia and problems giving any qualified guess due to their disorder. Both patients felt most comfortable using a combination of pointing and repeating when responding. The experimenter sitting next to the patient was watching his mouth movements and recorded the answer. The control participants felt most comfortable typing their responses on a keyboard.

In order to accustom the participants to the experimental setting and evaluate whether they were able to perform the experiment at all we conducted a number of training blocks containing one presentation of each stimulus in random order. The first part of the training contained only the congruent audiovisual stimuli. Both patients felt confident with the task after a single block. The next part of the training contained only the auditory stimuli. ML required three blocks and MK required two blocks before feeling confident about the task. The final part of the training contained only the visual stimuli. ML was certain that he was unable to identify any of the visual stimuli after two blocks whereas MK felt confident about the task after a single block. Performance in the training blocks was not perfect for either of the two patients. The healthy control participants were confident about the task after a single block for auditory, visual and audiovisual stimuli and their performance was perfect. After the training we proceeded to the actual experiment in which 10 repetitions of each of the auditory, visual, and audiovisual stimuli were presented in random order. The participants were instructed to look at the mouth and to report what they heard when they heard a voice (in auditory and audiovisual stimuli) but lip-read when there was no voice (in visual-only stimuli).

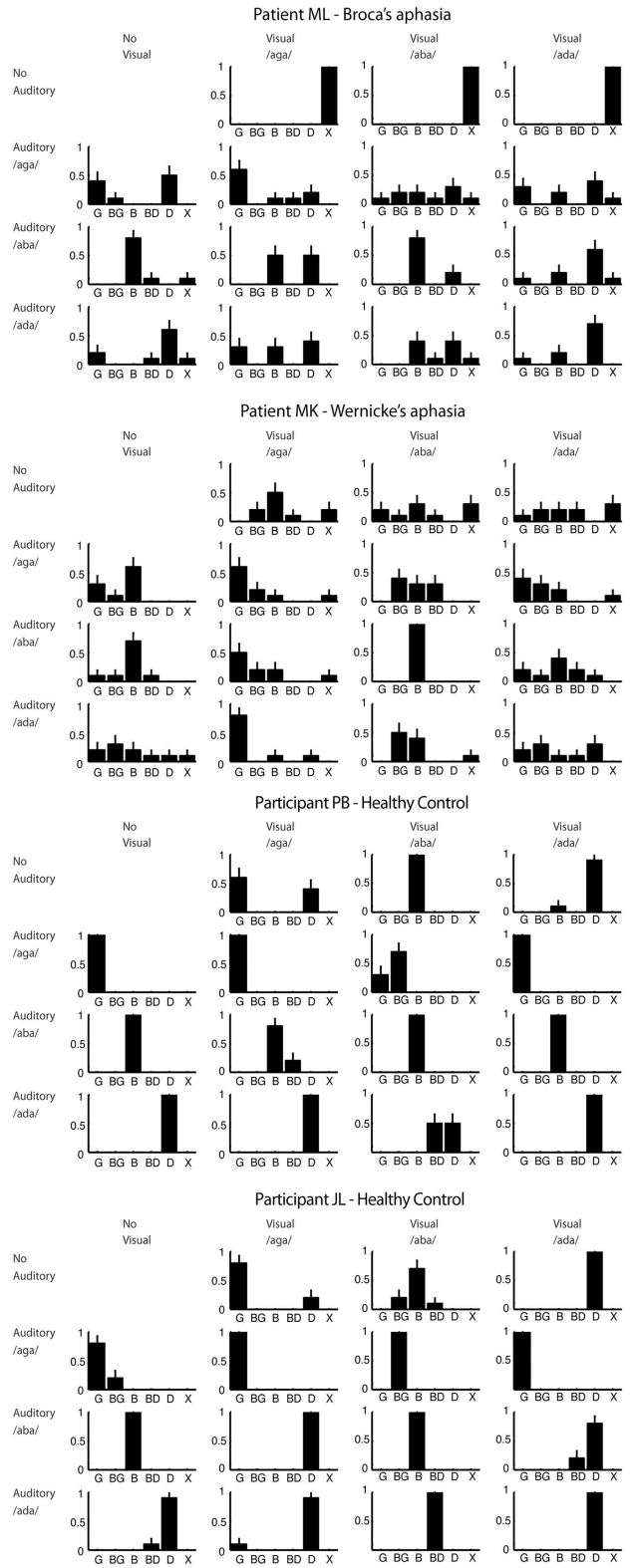


FIGURE 2 | Response proportions for all participants. Plots are arranged in rows and columns according to the visual and auditory component of the stimulus respectively. The response category "X" corresponds to no answer. Error bars show the multinomial standard deviation.

Results

The response proportions for all participants are shown in **Figure 2** from which it can be seen that both patients showed great variability in their responses compared to the healthy control participants. This could indicate that the patients found the task difficult in general.

Visual speech perception was particularly poor for the patients. Patient ML refused to venture a guess, which may reflect an inability to lip-read but could also reflect a lack of confidence in his own ability. Patient MK's lip-reading was also poor at 10% correct overall and poor even for /aba/ featuring a visually distinct bilabial closure. The percentage correct was 83% for each of the control participants showing that the articulations in the videos were sufficiently clear to be lip-read.

For incongruent audiovisual stimuli, in which the visual stimulus is a bilabial closure, the typical McGurk illusion is a combination illusion of hearing the bilabial closure followed by the auditory stimulus (e.g., auditory /ada/ + visual /aba/ = /abda/). This illusion was perceived by the two control participants, JL and PB respectively, in 100% and 60% of the trials including those stimuli. Only 20% of Patient ML's responses to the two incongruent audiovisual stimuli with a bilabial visual component fell in the combination response categories. For Patient MK, this number was 60%. This difference could indicate that audiovisual integration was weaker in ML for these stimuli. However, as the strength and type of McGurk illusion varies between healthy observers the difference observed here could also be due to normal variation between observers unrelated to patients' type of aphasia.

Incongruent audiovisual stimuli in which the acoustic stimulus is /b/ often create strong fusion and/or visual dominance illusions. For auditory /b/ with visual /g/ the typical McGurk illusion is a fusion illusion of hearing /d/ although a visual dominance illusion of hearing /g/ also occurs (MacDonald and McGurk, 1978; Andersen et al., 2009). Auditory /b/ with visual /d/ typically leads to a visual dominance illusion of hearing /d/ although /g/ responses have also been reported (MacDonald and McGurk, 1978). Somewhat surprisingly, whereas control participant PB did not experience these illusions for these stimuli at all, participant JL perceived them in 90% of the trials. This discrepancy is, however, not unusual as there is great inter-individual variability in the strength of the McGurk illusion (Magnotti and Beauchamp, 2014). For these two stimuli 40% of ML's responses fell in the visual dominance and fusion response categories. The same number for MK was 60%.

For audiovisual stimulus combinations of /aga/ and /ada/ (either visual /aga/ with auditory /ada/ or vice versa) observers often do not experience a McGurk illusion but respond according the auditory component of the stimulus (MacDonald and McGurk, 1978). We initially included these stimuli in the experiment for completeness although they arguably do not bare much evidence for or against our hypothesis. In fact, as the two healthy control participants did not experience the McGurk illusion for these stimuli, including them in a pooled analysis may actually lead to an under-estimation of the strength of the

McGurk illusion in the patients. We therefore conducted a pooled analysis omitting these stimuli.

In order to test our main hypothesis, whether ML experience the McGurk illusion, with some statistical power we first pool responses according to three stimulus types (incongruent audiovisual, auditory, and congruent audiovisual) and two response categories (correct and incorrect according to the auditory stimulus). If an observer is influenced by visual speech in the audiovisual conditions the proportion correct should be lower in the incongruent conditions compared to the congruent conditions. Additionally, we would expect the proportion correct in the auditory condition to be higher than in the incongruent condition and lower than in the congruent condition although this latter effect is sensitive to ceiling effects (Sumby and Pollack, 1954).

Figure 3 shows the proportion correct (according to the auditory stimulus) of the three stimulus types. Both patients showed a significantly higher proportion correct for congruent audiovisual stimuli than for incongruent audiovisual stimuli ($p < 10^{-5}$ for patient ML, $p < 10^{-7}$ for patient MK, one-sided Fisher's exact test). The proportion correct for auditory stimuli was also higher than for incongruent audiovisual stimuli ($p < 10^{-4}$ for patient ML, $p < 10^{-3}$ for patient MK, one-sided Fisher's exact test). This shows that both patients where influenced by visual speech. These comparisons were also highly significant ($p < 10^{-9}$ for all tests, one-sided Fisher's exact test) for the control participants. For patient MK the proportion correct for congruent audiovisual stimuli was also significantly higher than the proportion correct for auditory stimuli ($p < 0.04$, one-sided Fisher's exact test). This difference was not significant for patient ML. It was also not significant for the control participants, which is probably due to ceiling effects only.

Discussion

The sight of the talking face influenced speech perception significantly in Patient ML with Broca's aphasia. This was evident as a smaller proportion correct for the incongruent audiovisual stimuli relative to the auditory stimuli. This indicates that ML experienced a McGurk illusion and, hence, that an intact Broca's

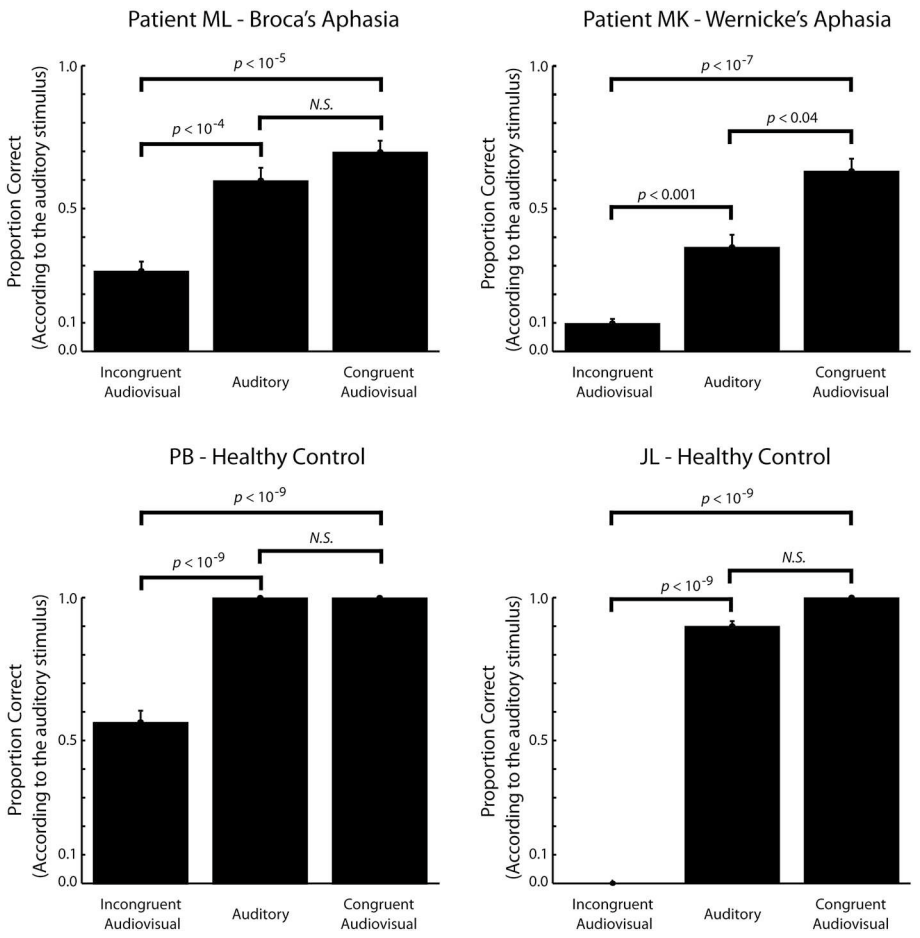


FIGURE 3 | The proportion correct (according to the auditory stimulus) as a function of stimulus type. For the incongruent audiovisual stimuli, a low proportion correct reflects a strong McGurk illusion. Only stimuli expected to elicit visual dominance, fusion or combination illusions are included in the incongruent audiovisual stimulus category (see text for details). Within subject comparisons are based on Fisher's exact one-sided test. Error bars show the binomial standard deviation.

area is not required for audiovisual integration of speech. A similar effect was found in Patient MK and the healthy control participants.

In addition to the McGurk illusion, patient MK with Wernicke's aphasia showed a congruency effect evident as significantly larger proportion correct for congruent audiovisual stimuli relative to auditory stimuli. Subject ML and the two healthy control participants did not show a similar effect. For the two healthy control participants this is due, at least partly, to a ceiling effect as they were nearly perfect in identifying the auditory stimuli leaving little room for improvement. In comparison, MK, showed very poor performance for auditory stimuli, which is in good agreement with his aphasia being receptive. For subject ML we cannot exclude a ceiling effect as, perhaps, his performance for this task is capped at a fairly low level due to his injury in general. On the other hand, we cannot exclude that this somehow reflects a more subtle effect of his injury to Broca's area more specifically.

The percentage correct in the lip-reading task was very low for both patients. Although poor lip-reading skills in aphasic patients have been reported previously (Youse et al., 2004; Klitsch, 2008; Hessler et al., 2012) this is surprising given the effect that visual speech had on auditory speech perception in both patients. This could imply a dissociation between visual and audiovisual speech, which could be due to conscious lip-reading taking place in a neural pathway, which is different from the one that affects auditory processing (Bernstein and Liebenthal, 2014). A similar dissociation was found in a study showing that point-light visual speech can cause a McGurk illusion even when observers are unaware of its speech-like nature and, hence, unable to lip-read (Rosenblum and Saldaña, 1996). This, to some degree, mimics how the patients seemed to perceive visual and audiovisual speech in the current study. Rosenblum and Saldaña (1996) also showed that static faces generally do not cause a McGurk effect even though they can be lip-read to some degree. Hence, where visual speech perception can be based on both static and dynamic information, only dynamic visual information is integrated with acoustic features. In support of this dissociation Calvert and Campbell (2003) found that lip-reading static faces activate different cortical areas than does lip-reading of dynamic features. As static natural images contain both configurational and featural facial information while point-light speech contains only configurational information another possible dissociation would be that visual speech perception can be based on both configurational and featural information while only featural information is integrated audio-visually. Both of these functional dissociations can be related to dual-stream anatomical models where the dorsal pathway relays mainly dynamic/featural information while the ventral pathway relays mainly static/configuration information (Bernstein and Liebenthal, 2014). Although this picture might not reflect the true complexity of the underlying neural structure (Bernstein and Liebenthal, 2014) it serves to show that it is not unlikely that lesions causing aphasia can influence visual and audiovisual speech perception differently.

Overall, performance in both patients was quite variable with many atypical responses. This variability matches that seen in previous studies of audiovisual speech perception in aphasic patients (Campbell et al., 1990; Youse et al., 2004; Klitsch, 2008; Hessler et al., 2012). These studies generally agree that aphasics, in general, do experience the McGurk illusion and thus integrate auditory and visual information in speech perception. They also agree that aphasics generally have poor lip-reading skills. Only few of them attempted a distinction between subtypes of aphasia and none have reached a conclusion on the specific role of Broca's aphasia.

A number of studies have previously examined audiovisual integration of speech in aphasic patients. Campbell et al. tested a patient that was diagnosed with aphasia after a cardiovascular accident (CVA) in the left MCA (Campbell et al., 1990). At the time of testing he was a fluent speaker experiencing that speech "sounded funny" indicating that his aphasia was mild and mainly receptive. He was also slightly dyslexic. This patient's performance was fairly poor (<50% correct) in a consonant identification task but the patient seemed to experience visual dominance and fusion illusions for incongruent audiovisual stimuli and Campbell et al. concluded that he was influenced by visual speech in his performance. This is in agreement with his lip-reading ability, which, for words, fell within the normal range.

Youse et al. tested a patient with mild anomic aphasia based on the Western Aphasia Battery, after a CVA affecting the territory of the MCA not unlike patient ML in the current study (Youse et al., 2004). Apparently, this patient showed a strong McGurk effect for auditory /bi/ with visual /gi/, which he perceived as /di/. Youse et al. did however note that this finding could be influenced by a strong response bias toward /di/ apparent in the auditory only condition. For auditory /gi/ with visual /bi/ they did not observe the typical combination illusion of perceiving /bgi/ but, in our opinion, a fairly strong visual dominance illusion of perceiving /bi/. Youse et al. concluded that the results did not show clear evidence of audiovisual integration in this patient. Notably, this patient's ability to lip-read was also very poor.

Klitsch (2008) investigated the McGurk effect in a group of six aphasic patients of which three were diagnosed with Broca's aphasia and the other three were diagnosed with Wernicke's aphasia. She found that the aphasic group experienced the McGurk illusion but did not distinguish between the types of aphasia. The proportion correct for lip-reading was 50% for this group given three response categories.

Hessler et al. (2012) investigated audiovisual speech perception in three aphasics and an age-matched control group. The three aphasics, were diagnosed as Wernicke's, anomic and mixed respectively. The mixed aphasic had suffered an ischaemic CVA in the left MCA, as our patient ML, and showed audiovisual interactions by giving more correct responses to congruent audiovisual speech than to auditory speech. This participant was tested on one incongruent combination of auditory and visual phonemes, auditory /p/ and visual /g/, and seemed to perceive the McGurk illusion in 61% of the experimental trials as measured by the proportion of incorrect responses. However, this should be compared to an error rate of 45% in the auditory condition—an error rate which was averaged across three different stimulus types, /k/, /p/, and /t/. Hessler et al. did not directly compare how well auditory /p/

was perceived when presented acoustically and when presented audiovisually, dubbed onto visual /k/. Hessler et al. concluded that none of the aphasic participants showed a preference for McGurk-type answers. The lip-reading performance for the three aphasic patients was, accordingly, poor ranging from 24 to 52% correct given only three response categories. For the control group, the corresponding range was 67–93%.

These studies by Hessler et al. and Klitsch suffered from very weak McGurk illusions (as low as 20% in Klitsch' study) even in the healthy control groups. Both studies ascribe this to a language effect specific to the Dutch language in which the studies were conducted. This makes it difficult to interpret whether the results in the aphasic group are representative.

Based on the studies described above, the poor lip-reading skills and atypical McGurk illusions that we found in Patient ML with Broca's aphasia seem to fall within the range generally seen in the aphasic population. Therefore, they cannot be ascribed to his lesion in Broca's area specifically. The variability seen across participants in these studies is not specific to aphasic patients but is also seen across studies of healthy subjects (Colin and Radeau, 2003). It is likely to be due, not only to variation across participants but also to variation across stimuli (Magnotti and Beauchamp, 2014). Hence, specific differences that are statistically significant are likely to be seen even between healthy participants and, to our knowledge, the reasons for this are unknown. Statistical comparisons of specific differences between aphasic patients or between patients and healthy controls will therefore be difficult to interpret without larger populations (Saalasti et al., 2011a,b; Meronen et al., 2013). Therefore, rather than conducting statistical analyses between participants, we limited our analysis to showing that audiovisual integration did take place in a patient with Broca's aphasia.

In accordance with our findings, Fridriksson et al. (2012) recently showed that speech production in Broca's aphasics can improve dramatically when they shadow audiovisual speech. Notably, this effect does not occur for auditory or visual speech. This shows that Broca's aphasics have some ability to integrate auditory and visual speech for use in speech production but the study does not address whether it also influences speech perception in general. As speech perception and production have been suggested to receive information from two, anatomically distinct, parallel streams (Hickok and Poeppel, 2007) the perceptual McGurk effect in Broca's aphasics may be due to a different mechanism than that studied by Fridriksson et al.

In summary, our findings show that speech perception in a patient with Broca's aphasia in influenced by the sight of the talking face as this patient experiences a McGurk illusion, which, although somewhat atypical, is little different from the McGurk effect seen in the aphasic population in general. This offers no confirmation of the hypothesis that Broca's area should be necessary for audiovisual integration of speech, contrary to Ramachandran et al.'s (1999) preliminary findings, and hence, no confirmation of the strong mirror neuron MT. Whether a lesion in Broca's area can have a more subtle effect on audiovisual integration, as it has on auditory speech perception, cannot be ruled out by the current results. Therefore, we do not consider our findings in disagreement with recent, more moderate versions of the MT in which articulatory representations have a subtle role in speech perception and are located in a distributed network of brain structures.

Acknowledgments

The authors wish to thank Tzvetelina S. Delfi for describing the CT-scans, Kasper Eskelund for help with running the experiment with patient MK and the participants who volunteered for the study. Also thanks to Sande/Klepp at Stavanger scan-support, and Fakutsi as always.

References

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., and Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Commun.* 51, 184–193. doi: 10.1016/j.specom.2008.07.004

Andersen, T. S., Tiippana, K., Lampinen, J., and Sams, M. (2001). "Modeling of audiovisual speech perception in noise," in *International Conference on Auditory–Visual Speech Processing (AVSP)* (Aalborg), 172–176.

Bernal, B., and Ardila, A. (2009). The role of the arcuate fasciculus in conduction aphasia. *Brain* 132, 2309–2316. doi: 10.1093/brain/awp206

Bernstein, L. E., and Liebenthal, E. (2014). Neural pathways for visual speech perception. *Front. Neurosci.* 8:386. doi: 10.3389/fnins.2014.00386

Binkofski, F., and Buccino, G. (2004). Motor functions of the Broca's region. *Brain Lang.* 89, 362–369. doi: 10.1016/S0093-934X(03)00358-4

Blumstein, S. E., Baker, E., and Goodglass, H. (1977). Phonological factors in auditory comprehension in aphasia. *Neuropsychologia* 15, 19–30. doi: 10.1016/0028-3932(77)90111-7

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat. Vis.* 10, 433–436.

Broca, M. P. (1861). Remarques sur le siége de la faculté du langage articulé, suivies d'une observation d'aphémie (Perte de la Parole). *Bulletins et Memoires de la Societe Anatomique de Paris* 36, 330–357.

Calvert, G. A., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* 15, 57–70. doi: 10.1162/089892903321107828

Campbell, R., Garwood, J., Franklin, S., Howard, D., Landis, T., and Regard, M. (1990). Neuropsychological studies of auditory-visual fusion illusions. Four case studies and their implications. *Neuropsychologia* 28, 787–802.

Colin, C., and Radeau, M. (2003). Les illusions McGurk dans la parole: 25 ans de recherches. *L'année Psychologique* 103, 497–542. doi: 10.3406/psy.2003.29649

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402. doi: 10.1046/j.0953-816x.2001.01874.x

Ferrari, P. F., Gallese, V., Rizzolatti, G., and Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur. J. Neurosci.* 17, 1703–1714. doi: 10.1046/j.1460-9568.2003.02601.x

Fridriksson, J., Hubbard, H. I., Hudspeth, S. G., Holland, A. L., Bonilha, L., Fromm, D., et al. (2012). Speech entrainment enables patients with Broca's aphasia to produce fluent speech. *Brain* 135, 3815–3829. doi: 10.1093/brain/aws301

Galantucci, B., Fowler, C. A., and Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychon. Bull. Rev.* 13, 361–377. doi: 10.3758/BF03193857

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119(Pt 2), 593–609. doi: 10.1093/brain/119.2.593

Hessler, D., Jonkers, R., and Bastiaanse, R. (2012). Processing of audiovisual stimuli in aphasic and non-brain-damaged listeners. *Aphasiology* 26, 83–102 doi: 10.1080/02687038.2011.608840

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Jääskeläinen, I. P. (2010). The role of speech production system in audiovisual speech perception. *Open Neuroimag. J.* 4, 30–36. doi: 10.2174/1874440001004020030

Kertesz, A. (1982). *Western Aphasia Battery*. New York, NY: Grune and Stratton.

Kleiner, M., Brainard, D., and Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception* 36 ECVP Abstract Supplement.

Klitsch, J. U. (2008). *Open Your Eyes and Listen Carefully*. Groningen: Dissertations in Linguistics (GRODIL).

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., et al. (2000). Handedness and hemispheric language dominance in healthy humans. *Brain* 123(Pt 12), 2512–2518. doi: 10.1093/brain/123.12.2512

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., and Gallese, V. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297, 846–848. doi: 10.1126/science.1070311

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist* 37, 148–167. doi: 10.1037/0003-066X.37.2.148

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lotto, A. J., Hickok, G. S., and Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends Cogn. Sci.* 13, 110–114. doi: 10.1016/j.tics.2008.11.008

MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. *Percept. Psychophys.* 24, 253–257. doi: 10.3758/BF03206096

Magnotti, J. F., and Beauchamp, M. S. (2014). The noisy encoding of disparity model of the McGurk effect. *Psychon. Bull. Rev*. doi: 10.3758/s13423-014-0722-2. [Epub ahead of print].

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Meronen, A., Tiippana, K., Westerholm, J., and Ahonen, T. (2013). Audiovisual speech perception in children with developmental language disorder in degraded listening conditions. *J. Speech Lang. Hear. Res.* 56, 211. doi: 10.1044/1092-4388(2012/11-0270)

Moineau, S., Dronkers, N. F., and Bates, E. (2005). Exploring the processing continuum of single-word comprehension in aphasia. *J. Speech Lang. Hear. Res.* 48, 884–896. doi: 10.1044/1092-4388(2005/061)

Möttönen, R., and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825 doi: 10.1523/JNEUROSCI.6018-08.2009

Möttönen, R., and Watkins, K. E. (2012). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology* 26, 1103–1118. doi: 10.1080/02687038.2011.619515

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage* 25, 333–338. doi: 10.1016/j.neuroimage.2004.12.001

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Ramachandran, V. S., Rogers-Ramachandran, D., Altschuler, E. L., and Wisdom, S. B. (1999). A test of Rizzolatti's theory of language evolution; McGurk effect and lip reading in Broca's aphasia? *Soc. Neurosci. Abstr.* 25:1636. Program No. 654.11.

Rauschecker, J. P., and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194. doi: 10.1016/S0166-2236(98)01260-0

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Rosenblum, L. D., and Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 22, 318–331.

Saalasti, S., Kätsyri, J., Tiippana, K., Laine-Hernandez, M., Wendt von, L., and Sams, M. (2011a). Audiovisual speech perception and eye gaze behavior of adults with asperger syndrome. *J. Autism Dev. Disord.* 42, 1606–1615. doi: 10.1007/s10803-011-1400-0

Saalasti, S., Tiippana, K., Kätsyri, J., and Sams, M. (2011b). The effect of visual spatial attention on audiovisual speech perception in adults with Asperger syndrome. *Exp. Brain Res.* 213, 283–290. doi: 10.1007/s00221-011-2751-7

Scott, S. K., McGettigan, C., and Eisner, F. (2009). A little more conversation, a little less action—candidate roles for the motor cortex in speech perception. *Nat. Rev. Neurosci.* 10, 295–302. doi: 10.1038/nrn2603

Sekiyama, K., and Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am.* 90, 1797–1805.

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89 doi: 10.1016/j.neuroimage.2004.11.006

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Watkins, K. E., Strafella, A. P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41, 989–994. doi: 10.1016/S0028-3932(02)00316-0

Watkins, K., and Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. *J. Cogn. Neurosci.* 16, 978–987. doi: 10.1162/0898929041502616

Wernicke, C. (1874). *Der Aphasische Symptomencomplex: Eine Psychologische Studie auf Anatomischer Basis*. Breslau: Max Cohn and Weigert.

Youse, K. M., Cienkowski, K. M., and Coelho, C. A. (2004). Auditory-visual speech perception in an adult with aphasia. *Brain Inj.* 18, 825–834. doi: 10.1080/02699000410001671784

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 May 2014
doi: 10.3389/fpsyg.2014.00416

How is the McGurk effect modulated by Cued Speech in deaf and hearing adults?

Clémence Bayard*, Cécile Colin and Jacqueline Leybaert

Center for Research in Cognition and Neurosciences, Université Libre de Bruxelles, Brussels, Belgium

Edited by:
Jean-Luc Schwartz, Centre National de la Recherche Scientifique, France

Reviewed by:
Bencie Woll, University College London, UK
Pascal Barone, Centre National de la Recherche Scientifique, France

****Correspondence:***
Clémence Bayard, Laboratoire Cognition Langage et Développement, Center for Research in Cognition and Neurosciences, Université Libre de Bruxelles, 50 Avenue Franklin Roosevelt – CP191, 1050 Brussels, Belgium
e-mail: clemence.bayard@ulb.ac.be

Speech perception for both hearing and deaf people involves an integrative process between auditory and lip-reading information. In order to disambiguate information from lips, manual cues from Cued Speech may be added. Cued Speech (CS) is a system of manual aids developed to help deaf people to clearly and completely understand speech visually (Cornett, 1967). Within this system, both labial and manual information, as lone input sources, remain ambiguous. Perceivers, therefore, have to combine both types of information in order to get one coherent percept. In this study, we examined how audio-visual (AV) integration is affected by the presence of manual cues and on which form of information (auditory, labial or manual) the CS receptors primarily rely. To address this issue, we designed a unique experiment that implemented the use of AV McGurk stimuli (audio /pa/ and lip-reading /ka/) which were produced with or without manual cues. The manual cue was congruent with either auditory information, lip information or the expected fusion. Participants were asked to repeat the perceived syllable aloud. Their responses were then classified into four categories: audio (when the response was /pa/), lip-reading (when the response was /ka/), fusion (when the response was /ta/) and other (when the response was something other than /pa/, /ka/ or /ta/). Data were collected from hearing impaired individuals who were experts in CS (all of which had either cochlear implants or binaural hearing aids; $N = 8$), hearing-individuals who were experts in CS ($N = 14$) and hearing-individuals who were completely naïve of CS ($N = 15$). Results confirmed that, like hearing-people, deaf people can merge auditory and lip-reading information into a single unified percept. Without manual cues, McGurk stimuli induced the same percentage of fusion responses in both groups. Results also suggest that manual cues can modify the AV integration and that their impact differs between hearing and deaf people.

Keywords: multimodal speech perception, Cued Speech, cochlear implant, deafness, audio-visual speech integration

INTRODUCTION

In face-to-face communication, speech perception is a multimodal process involving mainly auditory and visual (lip-reading) modalities (Sumby and Pollack, 1954; Grant and Seitz, 2000). Hearing-people merge auditory and visual information into a unified percept, a mechanism called audio-visual integration (AV integration). This merging of information has been demonstrated through the McGurk effect (McGurk and MacDonald, 1976), in which integration occurs even when auditory and visual modalities provide incongruent information. For example, the simultaneous presentation of the visual velar /ka/ and auditory bilabial /pa/ normally leads hearing-individuals to perceive the illusory *fusion* alveo-dental /ta/. The McGurk effect suggests that visual articulatory cues about place of articulation are integrated into the auditory percept which is then modified.

Presently, many children born deaf are fitted with cochlear implants (CI). This technology improves a child's ability to access auditory information. Studies have shown that deaf individuals (both adults and children) whom of which were fitted with CI's were able to integrate auditory and visual information, with better performance in the AV condition than in the audio condition (Erber, 1972; Tyler et al., 1977; Hack and Erber, 1982; Lachs et al., 2001; Geers et al., 2003; Bergeson et al., 2005; Desai et al., 2008). However, auditory information provided by the CI was degraded with respect to place of articulation, voicing and nasality (Dowell et al., 1982; Skinner et al., 1999; Kiefer et al., 2001). Therefore, participants fitted with a CI gave more importance to lip-read information in AV speech integration than did hearing participants (Schorr et al., 2005). In the case of incongruent auditory and visual information (McGurk stimuli), deaf participants (adults and children) gave more responses based on visual information, whereas hearing participants gave more integration responses or responses based on auditory information (Leybaert and Colin, 2007; Desai et al., 2008; Rouger et al., 2008; Huyse et al., 2013). However, the reliance on lip-reading information was flexible: when visual information was degraded, children with CI's relied less on visual information, and more on auditory information (Huyse et al., 2013). The AV integration is thus an adaptive process in which the respective weights of each modality depend on the level of uncertainty in auditory and visual signals.

Aside from lip-reading, Cued Speech could help deaf people overcome the uncertainty of auditory signals delivered by

the CI. Originally, the Cued Speech (CS) system was designed to help deaf people (without a CI) perceive speech through disambiguating the visual modality (Cornett, 1967). The CS system reduces the ambiguity related to lip-reading by making each of the phonological contrasts of oral language visible. Each syllable is uttered with a complementary gesture called a manual cue. CS was adapted to the French language in 1977, and is currently known as "Langue française Parlée Complétée." In French, the vowels are coded with five different hand placements near the face, and consonants are coded with eight hand-shapes (see **Figure 1**). Each manual cue can code several phonemes, but these phonemes differ in their labial image. Also, consonants and vowels sharing the same labial image are coded by different cues. Thus, the combination of visual information, provided by the articulatory labial movements and manual cues, allows deaf individuals to correctly perceive all syllables. Nicholls and Ling (1982) studied the benefits of CS on speech perception. They compared deaf children's speech perception with or without CS and showed that the addition of CS improves speech perception from 30 to 40% in a lip-reading-only condition to 80% with the addition of manual cues. Similar results were found with French CS (Périer et al., 1990). Exposure to CS contributes to the elaboration of phonological representations, hence improving abilities notably in rhyme judgments, rhyme generation, spelling production as well as reading (Charlier and Leybaert, 2000; Leybaert, 2000; LaSasso et al., 2003; Colin et al., 2007).

While the advantages of exposure to CS are well-recognized, the processing of the CS signal still remains unclear. Attina et al. (2004) were the first to examine the precise temporal organization of the CS production of syllables, words, and sentences. They found that manual cues naturally anticipate lip gestures, with a maximum duration of 200 ms before the onset of the corresponding acoustic signal. In a second study, the same authors showed a propensity in deaf people to anticipate manual cues over lip cues during CS perception. That is to say, deaf people extract phonological information when a manual cue is produced whether or not lip movements are completed. This phonological extraction has the effect of reducing the potential number of

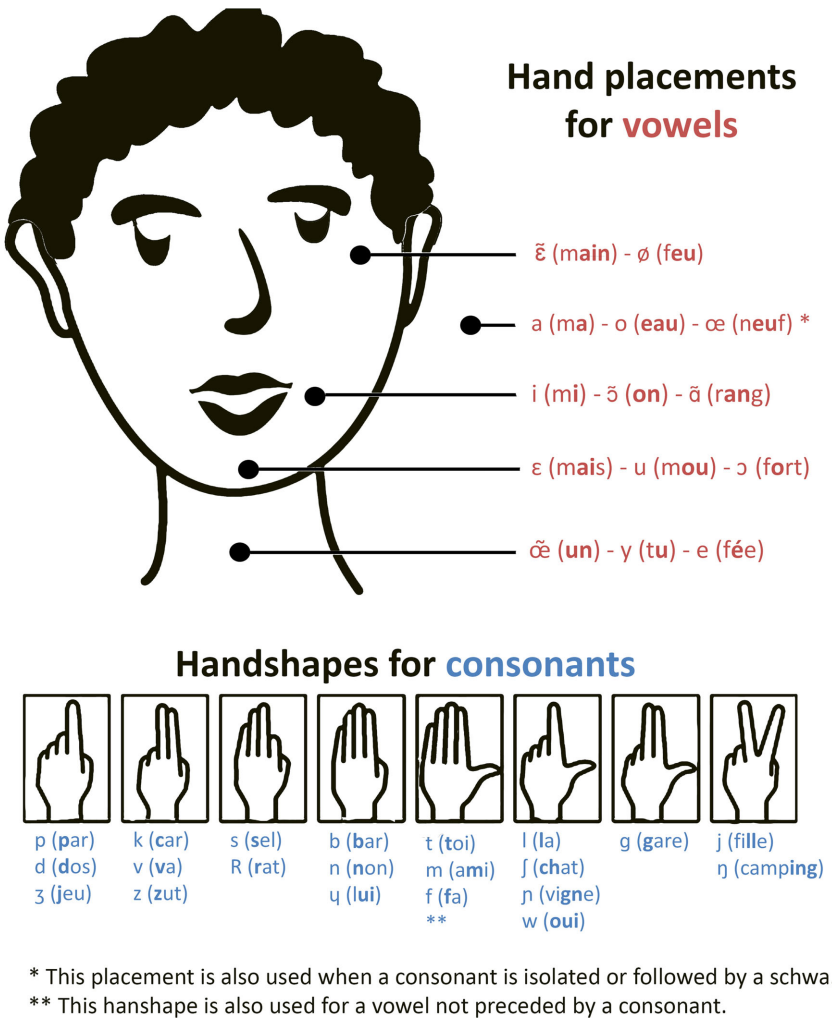


FIGURE 1 | Cues in French Cued Speech: hand-shapes for consonants and hand placements for vowels. Adapted from http://sourdsressources.wordpress.com.

syllables that could be perceived (Attina, 2005; Aboutabit, 2007; Troille et al., 2007; Troille, 2009). These results reverse the classic way of considering the CS system: manual cues, as opposed to labial information, could be the primary source of phonological information for deaf CS-users. Despite the fact that manual cues are artificial, they might constitute the main source of phonological information, and labial information would then be used to disambiguate this manual information.

Alegria and Lechat (2005) studied the integration of articulatory movements in CS perception. More precisely, they investigated the relative influence of labial and manual information on speech perception. Deaf children (mean age: 9 years, with normal intelligence and schooling) were split into two groups depending on their age of exposure to CS (early or late). They were asked to identify CV syllables uttered without manual cues (lip-reading alone) or with manual cues (Cued Speech). In the CS condition, lip movements and manual cues were either congruent (e.g., lip-reading /ka/ and hand-shape n°2, that codes /v, z, k/) or incongruent (e.g., lip-reading /ka/ and hand-shape n°1, that codes /d, p, ʒ/). Identification scores were better in the congruent and lip-reading alone condition than when syllables were presented with incongruent manual cues. In the incongruent condition, participants reported syllables coded with the same manual cues as the actual syllables. Between the different syllables coded by a matching manual cue, deaf participants selected the one that had less visible lip movements; that is, the one that was less inconsistent with lip information presented in the syllable stimuli. For example, the lip movements /ka/ with hand-shape n°1 (coding /d, p, ʒ /) was perceived as /da/ which is less visible on the lips than /pa/ and /ʒa/. This suggests an integrative process between lip and manual cue information. Moreover, deaf children who were exposed to CS early (prior to 2 years) integrated manual cue and lip-read information better than deaf children who were exposed to CS later (after 2 years). To conclude, when lip-read information and manual cues diverge, participants choose a compromise that is compatible with manual information and not incompatible with the lip-read one.

The goal of the present research was to examine how manual cue information is integrated in AV speech perception by deaf and hearing participants. We wondered whether (1) CS receptors combine auditory, lips and manual information to produce a unitary percept; (2) on which information (auditory, labial or manual) they primarily rely; and (3) how this integration is modulated by auditory status. To address these issues we designed the first experiment using audio-visual McGurk stimuli produced with manual cues. The manual cue was either congruent with auditory information, lip information or with the expected fusion. We examined whether or not these experimental conditions would impact the pattern of responses differently for deaf and hearing subjects.

MATERIALS AND METHODS

PARTICIPANTS

Thirty-seven adults participated in the study. They were split into three groups according to their auditory status and degree of CS expertise. The first group consisted of eight deaf CS users (mean age: 18 years), hereafter referred to as the CS-deaf group. Three of them had cochlear implants and five used binaural hearing aids. Seven had been exposed to CS from the age of two to three years and the remaining one from the age of 14 years (for more details see **Table 1**) The second group was comprised of 14 hearing CS users (mean age: 22 years), hereafter referred to as the CS-hearing group. Two of them had close relatives that were deaf; the rest were students in speech therapy and had participated in CS training sessions. The third group consisted of 15 hearing-individuals who had never been exposed to CS (mean age: 23 years), hereafter referred to as the control hearing group.

All participants were native French speakers with normal or corrected-to-normal vision and did not have any language or cognitive disorder. In order to assess CS knowledge level, a French CS reception test was administered to all participants (TERMO). Scores groups and participants are indicated in Appendix, **Table A1**. The experimental protocol was approved by the ethical committee of the Faculty of Psychological Science and Education (Université Libre de Bruxelles). All participants provided informed consent, indicating their agreement to participate in study. They were informed they had the option to withdraw from the study at any time.

EXPERIMENTAL MATERIAL

Stimuli

A female French speaker was videotaped while uttering CV syllables consisting of one of the /p, k, t/ consonants articulated with /a/ (**Figure 2**).

Congruent conditions

Two uni-modal and four multi-signal congruent conditions were created (see **Table 2**). They served as control conditions. Each stimulus from the congruent conditions was presented 6 times.

Incongruent conditions

Stimuli were also presented in incongruent conditions. Incongruent AV syllables were created by carefully combining audio files /pa/ with non-corresponding video files /ka/ and matching their onset. Four incongruent conditions were created which consisted of McGurk stimuli (audio/pa/ and lip-reading /ka/) presented with or without manual cues (see **Table 3**). Each stimulus from the incongruent condition was presented 6 times.

Table 1 | CS-deaf group characteristics.

Participants	Age (in years)	Age at diagnosis	Age at equipment (in years)	Age at CS exposure (in years)
1	17	At birth	Unknown	2
2	21	3 years	3	3
3	21	At birth	2	3
4	14	At birth	3	2
5	24	At birth	3	2
6*	21	At birth	5	2
7*	16	At birth	8	2
8*	17	2 years	16	14

*Indicates participants with cochlear implants.

Table 2 | Stimulus composition of congruent control conditions.

Conditions	Stimulus 1	Stimulus 2	Stimulus 3
Audio only	A /pa/	A /ta/	A /ka/
Lip-reading only	LR /pa/	LR /ta/	LR /ka/
Audio + CS cue	A /pa/ + CS cuecoding /**p**, d, ʒ/	A /ta/ + CS cuecoding /m, **t**, f/	A /ka/ + CS cuecoding /**k**, v, z/
Lip-reading + CS cue	LR /pa/ + CS cuecoding /**p**, d, ʒ/	LR /ta/ + CS cuecoding /m, **t**, f/	LR /ka/ + CS cuecoding /**k**, v, z/
Audio visual	A /pa + LR /pa/	A /ta/ + LR /ta/	A /ka/ + LR /ka/
AV + CS cue	A /pa/ + LR /pa/ + CScue coding /**p**, d, ʒ/	/	/

Because each CS cue codes several phonemes, the phoneme congruent with auditory information, or lip-reading information is indicated in bold.

Table 3 | The composition of McGurk stimuli in incongruent conditions.

	Auditory info.	Lip reading info.	Manual cue info.
Baseline condition	pa	ka	/
Audio condition	pa	ka	**pa**, da, ʒa (congruent with auditory information)
Lip-reading condition	pa	ka	**ka**, va, za (congruent with lip read information)
Fusion condition	pa	ka	ma, **ta**, fa (congruent with the expected fusion)

Because each CS cue codes several phonemes, the phoneme congruent with auditory information, or lip-read information, or the expected fusion is indicated in bold.

FIGURE 2 | Stimulus sample. Video frame of lip-reading with congruent cue condition **(A)**, of audio only condition **(B)**, of audio with congruent cue condition **(C)**.

PROCEDURE

The experiment took place in a quiet room. Videos were displayed on a 17.3 inch monitor on a black background at eye level and at 70 cm from the participant's head. The audio track was presented at 65 dB SPL (deaf participants used their hearing-aids during the experiment). On each trial, participants saw a speaker's video (duration 1000 ms; see **Figure 2**). They were then asked to repeat aloud the perceived syllable. Their answers were transcribed by the experimenter. The experiment consisted of 120 items (16×6 congruent stimuli and 4×6 incongruent stimuli) presented in two blocks of 60 items. In each block, all conditions were mixed. Before starting, participants were shown five training items. The total duration of the experiment was approximately 30 min.

RESULTS

CONGRUENT CONDITIONS

As the groups were small ($N < 15$), we used non-parametric tests. In the congruent condition, we wanted to compare participants according to two criteria: auditory status (hearing vs. deaf) and CS abilities (CS users vs. non-CS users). Mann-Whitney tests were used to compare hearing (CS and non-CS together) with deaf groups and to compare CS users (deaf and hearing together) with the control group.

Audio conditions (with or without CS cue)

As illustrated in **Table 4**, in the Audio-Only condition, deaf and hearing-individuals had the same percentage of correct responses for the stimulus /pa/ ($U = 91$; $p = 0.184$). As it appeared that the standard deviation for the deaf group (18.2) was much higher than that of the hearing group, we analyzed individual scores of the deaf participants. Participant 2 was the only one to have a score under 83%; he obtained only 17% of correct responses. As confirmed by TERMO scores (**Table 1**), despite his binaural hearing aids, participant 2 had a low level of auditory recovery. When data were re-analyzed without this atypical participant, the outcome remained unchanged: Deaf and hearing-individuals had the same percentage of correct responses for the stimulus /pa/ ($U = 91$; $p = 0.373$). However, the CS-deaf group had more difficulty than the two hearing groups in identifying stimuli /ta/ ($U = 29$; $p < 0.005$) and /ka/ ($U = 43.50$; $p < 0.005$). Compared to the Audio-Only condition, the addition of cues improved the percentages of correct answers for the CS-deaf group, nonetheless the hearing groups still had more correct responses for /pa/ ($U = 73.5$; $p < 0.05$), /ta/ ($U = 31.5$; $p < 0.001$) and /ka/ ($U = 87$; $p < 0.01$)

Lip-reading conditions (with or without CS cue)

In the Lip-reading-Only condition, both deaf and hearing participants had similar percentages of correct responses for /pa/ ($U = 77$; $p = 0.068$), /ta/ ($U = 157$; $p = 0.37$) and /ka/ ($U = 170.5$; $p = 0.173$). The addition of cues, in comparison with the Lip-reading-Only condition, increased the percentages of correct

Table 4 | Mean percentages of correct responses for all groups in Audio-Only and Audio + CS cue conditions.

	CS-deaf		CS-hearing		Control hearing	
	Audio only cond.	Audio + CS cue cond.	Audio only cond.	Audio + CS cue cond.	Audio only cond.	Audio + CS cue cond.
/pa/	85 (18.2)	93 (12.5)	100 (0)	98 (2.4)	98 (2.1)	95 (7.1)
/ta/	62 (21.9)	70 (23.9)	100 (0)	98 (0)	100 (0)	100 (0)
/ka/	59 (29.2)	93 (9.4)	100 (0)	100 (0)	100 (0)	100 (0)

Standard deviations are indicated in parentheses.

Table 5 | Mean percentages of correct responses for all groups in Lip-reading-Only and Lip-reading + CS cue conditions.

	CS-deaf		CS-hearing		Control hearing	
	Lip-reading-Only cond.	Lip-reading + CS cue cond.	Lip-reading-Only cond.	Lip-reading + CS cue cond.	Lip-reading-Only cond.	Lip-reading + CS cue cond.
/pa/	68 (18.8)	100 (0)	71 (18.7)	91 (9.9)	91 (10.7)	77 (17.8)
/ta/	52 (27.1)	85 (18.2)	38 (27.8)	69 (36.9)	46 (24)	38 (24.4)
/ka/	22 (14.6)	89 (15.6)	8 (11.0)	69 (22.9)	14 (13.5)	52 (24.9)

Standard deviations are indicated in parentheses.

answers for CS users (deaf and hearing). CS users had better responses than control participants for /pa/ ($U = 98$; $p < 0.05$), /ta/ ($U = 82.5$; $p < 0.01$), and /ka/ ($U = 98.5$; $p < 0.05$). Percentages of correct responses for each group are shown in **Table 5**.

Audio with Lip-reading conditions (with or without CS cue)

As illustrated in **Table 6**, deaf and hearing-individuals obtained 100% of correct responses for the AV stimulus /pa/. However, the CS-deaf group had more difficulty than either of the two hearing groups in identifying AV stimuli /ta/ ($U = 43.5$; $p < 0.01$) and /ka/ ($U = 43.5$; $p < 0.01$). Deaf participants did not obtain 100% of correct responses for stimuli /ta/ and /ka/, because both the audio and visual information were difficult to identify (audio /ta/ 62%, audio /ka/ 59%, lip-reading /ta/ 52% and lip-reading /ka/ 22%; **Tables 4**, **5**).

When all information (auditory, labial and manual) were presented, participants had the same percentage of correct responses for /pa/.

INCONGRUENT CONDITIONS

Participant responses were classified into four categories: audio (when the response was /pa/), lip-reading (when the response was /ka/), fusion (when the response was /ta/) and other. In the baseline condition, we used Mann-Whitney tests to compare hearing (CS and non-CS together) with deaf groups. In each group, the Wilcoxon test was used to compare response patterns between baseline and other experimental conditions.

McGurk—Baseline condition (audio /pa/ + lip-reading /ka/)

As illustrated in **Table 7**, deaf and hearing-individuals had the same percentages of fusion ($p = 0.39$) and auditory ($p = 0.18$) responses.

Table 6 | Mean percentages of correct responses for all groups in Audio + Lip-reading (LR) and Audio + LR + CS cue conditions.

	CS-deaf	CS-hearing	Control hearing
Audio /pa/ + LR /pa/	100 (0)	100 (0)	100 (0)
Audio /ta/ + LR /ta/	64 (27.1)	100 (0)	100 (0)
Audio /ka/ + LR /ka/	62 (26.0)	100 (0)	100 (0)
Audio /pa/ + LR /pa/ + CS /pa/	100 (0)	100 (0)	100 (0)

Standard deviations are indicated in parentheses.

McGurk—Audio condition (audio /pa/ + lip-reading /ka/ + CS cue coding /p,d,ʒ/)

Response patterns for each group in the McGurk-audio condition are shown in **Table 7**. Compared to the baseline condition, the addition of the /p, d, ʒ/ cue reduced the percentage of fusion responses in the CS-deaf group ($p = 0.03$) in favor of other responses congruent with cue information (60% of other responses: 38% of /da/ and 19% of /ʒa/). In the CS-hearing group, the addition of cue n°1 reduced the percentage of fusion responses ($p = 0.001$) and increased auditory responses from 17% to 60% ($p = 0.003$). In the Control hearing group, the addition of the cue had no effect on the response pattern.

McGurk—Lip-reading condition (audio /pa/ + lip-reading /ka/ + CS cue coding /k,v,z/)

As illustrated in **Table 7**, the addition of the cue coding /k, v, z/ in the CS-deaf group, reduced the percentage of fusion responses ($p = 0.02$) and increased the percentage of lip-reading responses ($p = 0.03$), in comparison with the baseline condition. In addition, some participants responded with the alternative, /za/, which was congruent with cue information. In the CS-hearing group, the addition of cue n°2 also decreased fusion responses

Table 7 | Mean percentages of each kind of response (audio, lip-reading, fusion and other) for all groups in incongruent conditions.

	CS-deaf	CS-hearing	Control hearing
McGurk—Baseline condition (audio /pa/ + lip-reading /ka/)			
Resp. audio /pa/	8 (14.6)	17 (20.5)	27 (28.9)
Resp. lip-reading /ka/	2 (3.6)	1 (2.4)	1 (2.1)
Resp. fusion /ta/	81 (24)	78 (20.7)	70 (29.3)
Other response	9 (10.4)	2 (4.3)	2 (2.1)
McGurk—Audio condition (audio /pa/ + lip-reading /ka/ + CS cue coding /p,d,ʒ/)			
Resp. audio /pa/	**18 (19.8)**	**60 (25)**	**37 (34.8)**
Resp. lip-reading /ka/	2 (3.6)	0 (0)	1 (2.1)
Resp. fusion /ta/	20 (27.1)	21 (22.5)	57 (32.9)
Other response	60 (31.2)	18 (21.5)	5 (5.8)
McGurk—Lip-reading condition (audio /pa/ + lip-reading /ka/ + CS cue coding /k,v,z/)			
Resp. audio /pa/	2 (3.6)	20 (21.1)	35 (33.4)
Resp. lip-reading /ka/	**60 (32.8)**	**40 (27.4)**	**2 (3.9)**
Resp. fusion /ta/	25 (22.9)	33 (24.1)	61 (30.4)
Other response	13 (18.7)	6 (7.9)	2 (2.1)
McGurk—Fusion condition (audio /pa/ + lip-reading /ka/ + CS cue coding /m,t,f/)			
Resp. audio /pa/	0 (0)	16 (23.7)	35 (33.8)
Resp. lip-reading /ka/	0 (0)	0 (0)	1 (2.1)
Resp. fusion /ta/	**91 (10.4)**	**75 (28.6)**	**61 (31.1)**
Other response	9 (10.4)	9 (13.8)	3 (3.9)

Standard deviations are indicated in parentheses. Audio, lip-reading or fusion response congruent with CS cue information are indicated in bold.

($p = 0.002$) and increased lip-reading responses ($p = 0.003$). In the Control hearing group, the addition of cue had no effect on the response pattern.

McGurk—Fusion condition (audio /pa/ + lip-reading /ka/ + CS cue coding /m,t,f/)

In all groups, the addition of the cue coding /m, t, f/ had no effect on response patterns (see **Table 7**). There was no increase of fusion responses when compared to the baseline condition.

DISCUSSION

The goal of the present study was to examine how manual cue information is integrated in AV speech perception. We examined whether CS receivers can combine auditory, lip and manual information to produce a unitary percept. We expected that CS would modulate the respective weights of lip-read and auditory information differently, depending on auditory status.

CUED SPEECH BENEFIT

The present data confirmed previous results (Nicholls and Ling, 1982; Périer et al., 1990) indicating that the addition of congruent cues to lip-read information improved performance in CS perception for CS users (both deaf and hearing). In the CS-deaf group, the percentage of correct answers rose respectively from 47.3% in the Lip-reading-Only condition to 91.3% in the Lip-reading with Manual Cue condition, whereas it increased from 39 to 76.3% in the CS-hearing group (see **Table 5**). CS is therefore an efficient system to aid deaf people in perceiving speech visually. Note that for the CS-deaf group, manual cues with audio information also showed an improvement in perception. Indeed, the percentage of correct responses increased from 68.7% in the Audio-Only condition to 85.3% in the Audio with Manual Cue condition (see **Table 4**).

In contrast, the addition of cues decreased performance for the control group. It seems as though the CS cue served as a distractor for this group causing a disruption in responses. Their attention could have been drawn to the hand gesture, resulting in less focus on lip-read information. Compared to the Lip-reading-Only condition, the addition of cues decreased their percentages of correct responses, despite showing no significant effect. Furthermore, in the McGurk conditions with manual cues, the presence of hand information possibly unbound audio and visual information. Being more attracted to irrelevant hand information than by lip information, participants tended to not integrate AV information, resulting in fewer fusion responses and favoring auditory responses.

AUDIO-VISUAL SPEECH INTEGRATION IN DEAF

Our results showed that deaf people with cochlear implants or binaural hearing aids can merge auditory and lip-reading information into a unified percept just as hearing-individuals do. In the baseline condition (audio /pa/ + lip-reading /ka/), percentages of fusion responses were high and similar for both hearing and deaf groups (74 and 81% respectively, **Table 7**). Contrary to previous studies (Leybaert and Colin, 2007; Desai et al., 2008; Rouger et al., 2008), deaf individuals did not tend to report more responses based on visual information than hearing-participants. One explanation might be that deaf and hearing-individuals both exhibited comparable levels of performance in uni-modal conditions: percentages for identification of the auditory syllable /pa/ and the lip-reading syllable /ka/ did not differ between neither deaf nor hearing groups.

MANUAL CUE EFFECT ON AUDIO-VISUAL SPEECH INTEGRATION

In the case of incongruent auditory and visual information (audio /pa/ and lip-reading /ka/), the addition of manual cues that were incongruent with the expected fusion response impacted the pattern of responses. For both deaf- and hearing-CS users, the proportion of fusion responses decreased. The CS system therefore has an effect on AV integration processes. In the case of congruency between manual cues and expected fusion, the CS system supports illusory perception. However, for all groups the percentage of fusion did not increase. One explanation might be that the proportion of fusion responses in the baseline condition was already fairly high in deaf and hearing groups (81 and 74%, respectively **Table 7**).

Whereas manual cues decreased fusion responses in both hearing- and deaf-CS users, their effect on other responses depended on auditory status. Indeed, the addition of manual cues congruent with auditory information (but not with lip-read information) increased only audio responses for /pa/ in the CS-hearing group but not in the CS-deaf group. In this latter group, fusion responses decreased in favor of other responses, congruent

with the manual cue coding /p, d, ʒ/ (i.e., response /da/ or / ʒa/). Thus, despite their good performance in the Audio-Only condition (85%), CS-deaf users seemed more confident with visual information (such as lip-read or manual cues). They were unable ignore lip-read information and relied more heavily on such information than on auditory.

The addition of manual cues congruent with lip-read information increased lip-reading responses in both groups. These results suggest that deaf- and hearing-CS users are capable of ignoring auditory information when such information is contradicted by lip-reading or manual cues. As the CS system is not necessarily used with auditory information, ignoring auditory information could be easier.

AUDITORY STATUS EFFECT OR AUDITORY ABILITIES EFFECT?

Deaf-CS users' multimodal speech perceptions differ from that of hearing CS-users. Our results have shown that the addition of manual cues congruent with auditory information impacts the speech perception of deaf and hearing-individuals differently. Perception for deaf individuals relies more on visual information (lip-reading and manual cues); whereas perception in hearing-CS users relies more on auditory information. This suggests that the processing of CS information is modulated by auditory status. We have envisioned two speech perception models in order to explain these results. As it is illustrated in **Figure 3A**, hearing-CS receptors integrate auditory and labial information first, before determining whether manual cues are helpful in assembling a coherent percept. While manual cues might precede labial and auditory stimuli (Attina et al., 2004), hearing-individuals are more prone to ignore manual information and give more auditory responses in lieu of incongruent AV stimuli. CS perception remains less natural for hearing-individuals than for deaf. In the second model (see **Figure 3B**), deaf-CS receptors first integrate manual and lip information before taking auditory information into account. Thus, deaf-CS users cannot ignore manual information, resulting in less auditory responses. However, in our experiment, the deaf-CS user group was too small of a sample to be split into two groups according to the participants' auditory recuperation. We were therefore not able to examine the effect of auditory recuperation on the nature of integration processes. Auditory status and auditory abilities were thus confounded, which renders our interpretation fragile.

Therefore, in a new study (Bayard et al., in preparation), we investigated whether auditory status or auditory abilities impact audio-lip-read-manual integration in speech perception by testing a larger sample of deaf individuals whom of which were fitted with cochlear implants. Our first collection of data suggests an effect of auditory ability. Deaf individuals with good auditory ability had the same pattern response as their hearing-counterparts. Thus, for hearing- and deaf individuals with good auditory speech perception abilities, speech perception may first involve an integration between auditory and lip-read information. The merged percept then could be impacted by manual information when such information is delivered (**Figure 3A**). For deaf individuals with low auditory ability, labial and manual information could be initially merged, and auditory information would be taken into account subsequently (**Figure 3B**).

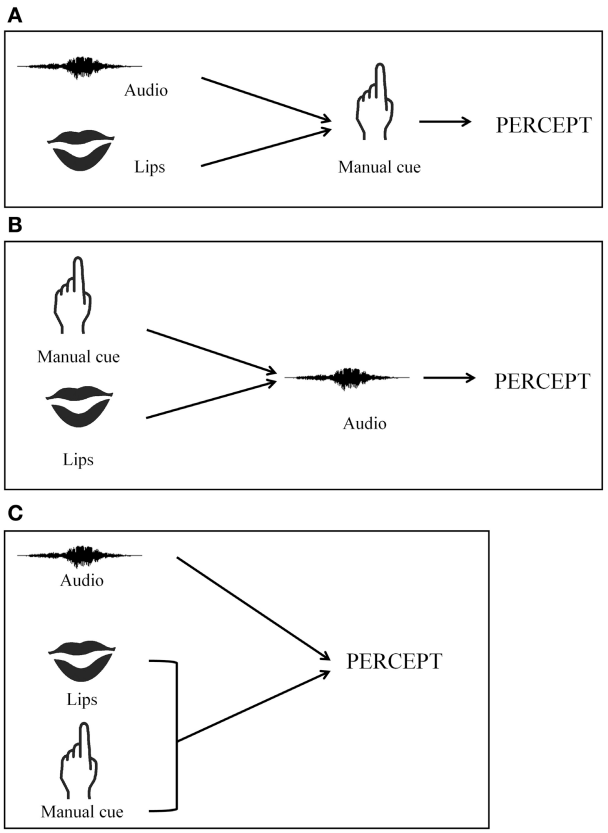


FIGURE 3 | CS perception models. (A) Sequential model with late integration of manual cue; **(B)** Sequential model with early integration of manual cue; **(C)** Simultaneous model with early integration of manual cue.

A number of other studies have revealed an impact of CI proficiency on AV speech integration. For example, Landry et al. (2012), compared three groups in a lip-reading task: proficient CI group, non-proficient CI group and normally-hearing group. Participants had to report visual speech stimulus presented in four conditions: visual only condition, AV speech condition, AV white noise condition, and AV reverse speech condition. Participants were informed that all auditory inputs were incongruent with the visual stimulus. Results showed that the presentation of auditory speech stimuli significantly impaired lip-reading performance only in proficient CI users and the normally-hearing group. Non-proficient CI users were not affected by auditory distractors, suggesting that such distraction was ignored due to their poor auditory ability. Huyse et al. (2013) showed that patterns of auditory, visual, and fusion responses to McGurk audio-visual stimuli are relative to CI proficiency. CI children who are AO− seemed to rely more on vision and CI children who are AO+ seemed to rely more on auditory information. Although these studies analyzed AV perception without cues, they reinforce our proposition that we should distinguish AO+ and AO− profiles in future studies of speech perception in participants with CI and CS.

INTEGRATION OF THE CS COMPONENT IN SPEECH PERCEPTION MODELS

Many AV integration studies on hearing-individuals have attempted to determine how and when integration takes place. More specifically, the issue of whether integration is early (before phonetic categorization) or late (after phonetic categorization) has been a topic of empirical and theoretical research. A number of speech perception and AV integration models have been proposed. Among such designs, the "Fuzzy logical model of perception" (FLMP; Massaro, 1987) postulates the existence of two stages in AV speech perception. The first stage is uni-modal processing. Auditory and visual features are assessed and compared to prototypes stored in memory. Comparison is based on a continued value scale and is independent in each modality. The second stage is bi-modal. Values of each feature are integrated in order to determine the degree of global adequacy of sensory input with each prototype in memory. The prototype that is the most consistent with the features extracted during the uni-modal assessment will be the percept heard. One important issue in this model is the fact that the influence of each source of information depends on its ambiguity. The more ambiguous the source, the less it influences perception. In addition, according to FLMP, all individuals integrate AV information optimally. In this way, all differences in the percept have to be explained by differences within the initial, uni-modal, stage.

The "Weight fuzzy logical model of perception" (WFLMP) is an interesting adaptation of FLMP (Schwartz, 2010). In WFLMP, inter-individual differences are taken into account. For each individual, specific weights may be allocated to each modality (visual and auditory). In WFLMP, differences in percept could be explained both by differences in uni-modal perception as well as by differences in integrative processing. As previous studies on speech perception in deaf-CI users have shown inter-individual differences (Landry et al., 2012; Huyse et al., 2013), the WFLMP seems to be more adapted than the FLMP in explaining such differences in perception. Recently, Huyse et al. (2013) conducted a study on speech perception in CI users and normally-hearing children. They tested the robustness of bias toward the visual modality in McGurk stimuli perception in CI users. For that reason, they designed an experiment in which the performances were compared in a "visual clear" condition and a "visual reduction" condition, in which the visual speech cues were degraded. Results showed that "visual reduction" had increased the number of auditory-based responses to McGurk stimuli, in normally-hearing as well as CI children (whose perception is generally dominated by vision). The authors used both FMLP and WFLMP to determine whether the differences in response patterns between "visual reduction" and "visual clear" conditions occurred at the uni-modal processing stage or at the integration stage. The FLMP model better fits the data in the "visual reduction" condition when an additional weight is applied to the auditory modality. The degradation of visual information seems to have an impact on speech perception not only at the uni-modal stage of processing but at the integrative processing level, as well. Thus, WFLMP seems to be a relevant model to explain AV speech perception in CI-users.

In the context of CI + CS perception, a third source of information is added: manual cue information. How is manual information processed in the WFMLP framework? We foresee three possibilities. According to a *first hypothesis*, the two types of visual information (manual cue and lip-read information) are processed in parallel and constitute the uni-modal, visual signal (**Figure 3C**). The influence of visual information (labial and manual) could be more important in both the uni-modal and integration stages of processing, in comparison to what occurs in classical AV integration. According to *the second hypothesis*, AV integration occurs as Schwartz described in WFLMP, and the manual cue information is merged with the AV percept later in integrative processing (**Figure 3A**). According to a *third hypothesis*, the labial- and manual-visual information are merged first, and auditory information is taken into account later (**Figure 3B**).

Currently, our studies have not allowed us to choose between these three hypotheses. It is clear that manual cue could impact AV integration. However, our behavioral data are not sufficient to determine whether this impact occurs early (as in the first hypothesis) or later (as in the second hypothesis). Furthermore, we have learned that deaf participants are capable of ignoring auditory cues, whereas they cannot ignore labial or manual information. Thus, for future studies, we aim to analyze more precisely the effect of auditory efficiency on speech perception, using data to confront our hypotheses.

In natural speech (without CS), humans speak and spontaneously produce gestures to support what they are saying. Analysis of speech and symbolic gesture production in adults suggest that both "are coded as a unique signal by a unique communication system" (Bernadis and Gentilucci, 2006). In addition, gestures play a crucial role in language development and a co-development of speech and signs exists (for a review see Capone and McGregor, 2004). Thus gesturing seems to be a genuine component of multi-modal communication. CS cues are created specifically for communication. Due to this privileged link between gestures and language, it is probable that these cues are naturally integrated into multi-modal communication. As shown by our data, it is difficult to ignore information provided by a cue.

CONCLUSION

Speech perception is a multimodal process in which different kinds of information are likely to be merged: naturally and relevant information (provided by lip-reading and audition), naturally but irrelevant information (like in audio-aerotactile integration), or non-natural but relevant information (such as CS cues).

Findings from our work also suggest that the integration of different types of information (e.g., audition, lip-reading, manual cues) related to a common source (i.e., the production of a speech signal) is a flexible process that depends on the informational content from the different sources of information, as well as on the auditory status and hearing proficiency of the participants.

ACKNOWLEDGMENTS

We thank Marie Devautour for her help in collecting data, Carol LaSasso and Jeromy Hrabovecky for their comments, suggestions and corrections of previous versions of this manuscript. This

work was financially supported by grant 2.4539.11 from the Fonds National de la Recherche Scientifique (FRS-FNRS, Belgium) to Jacqueline Leybaert and Cécile Colin, Clémence Bayard is funded by a Mini Arc award (ULB) for her PhD Thesis.

REFERENCES

Aboutabit, N. (2007). *Reconnaissance de la Langue Française Parlée Complétée (LPC): Décodage Phonétique des Gestes Main-Lèvre*. Ph.D. dissertion, Institut National Polytechnique de Grenoble, Grenoble.

Alegria, J., and Lechat, J. (2005). Phonological processing in deaf children: when Lip-reading and Cues are incongruent. *J. Deaf Stud. Deaf Educ.* 10, 122–133. doi: 10.1093/deafed/eni013

Attina, V. (2005). *La Langue Française Parlée Complétée: Production et Perception*. Ph.D. dissertion, Institut National Polytechnique de Grenoble, Grenoble.

Attina, V., Beautemps, D., Cathiard, M. A., and Odisio, M. (2004). A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthetizer. *Speech Commun.* 44, 197–214. doi: 10.1016/j.specom.2004.10.013

Bergeson, T. R., Pisoni, D. B., and Davis, R. A. O. (2005). Development of audiovisual comprehension skills in prelingually deaf children with cochlear implants. *Ear Hear.* 26, 149–164. doi: 10.1097/00003446-200504000-00004

Bernadis, P., and Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia* 44, 108–190.

Capone, N. C., and McGregor, K. K. (2004). Gesture development: a review for clinical and research practices. *J. Speech Lang. Hear. Res.* 47, 173–186. doi: 10.1044/1092-4388(2004/015)

Charlier, B., and Leybaert, J. (2000). The rhyming skills of deaf children educated with phonetically augmented speechreading. *Q. J. Exp. Psychol.* 53A, 349–375. doi: 10.1080/713755898

Colin, S., Magnan, A., Ecalle, J., and Leybaert, J. (2007). Relation between deaf children's phonological skills in kindergarten and word recognition performance in first grade. *J. Child Psychol. Psychiatry* 48, 139–146. doi: 10.1111/j.1469-7610.2006.01700.x

Cornett, R. O. (1967). Cued speech. *Am. Ann. Deaf* 112, 3–13.

Desai, S., Stickney, G., and Zeng, F.-G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *J. Acoust. Soc. Am.* 123, 428–440. doi: 10.1121/1.2816573

Dowell, R. C., Martin, L. F. A., Tong, Y. C., Clark, G. M., Seligman, P. M., and Patrick, J. F. (1982). A 12-consonant confusion study on a multiple-channel cochlear implant patient. *J. Speech Hear. Res.* 25, 509–516.

Erber, N. P. (1972). Auditory, visual and auditory visual recognition of consonants by children with normal and impaired hearing. *J. Speech Hear. Res.* 15, 407–412.

Geers, A. E., Nicholas, J. G., and Sedey, A. L. (2003). Language skills of children with early cochlear implantation. *Ear Hear.* 24, 46S–58S. doi: 10.1097/01.AUD.0000051689.57380.1B

Grant, K. W., and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208. doi: 10.1121/1.1288668

Hack, Z. C., and Erber, N. P. (1982). Auditory, visual and auditory-visual perception of vowels by hearing impaired children. *J. Speech Hear. Res.* 27, 100–107.

Huyse, A., Berthommier, F., and Leybaert, J. (2013). Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children. *Ear Hear.* 34, 110–121. doi: 10.1097/AUD.0b013e3182670993

Kiefer, J., Hohl, S., Stürzebecher, E., Pfennigdorff, T., and Gstöettner, W. (2001). Comparison of speech recognition with different speech coding strategies (SPEAK, CIS, and ACE) and their relationship to telemetric measures of compound action potentials in the nucleus CI 24M cochlear implant system. *Int. J. Audiol.* 40, 32–42. doi: 10.3109/00206090109073098

Lachs, L., Pisoni, D. B., and Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report. *Ear Hear.* 22, 236–251. doi: 10.1097/00003446-200106000-00007

Landry, S., Bacon, B. A., Leybaert, J., Gagné, J.-P., and Champoux, F. (2012). Audiovisual segregation in cochlear implant users. *PLoS ONE* 7:e33113. doi: 10.1371/journal.pone.0033113

LaSasso, C., Crain, K. L., and Leybaert, J. (2003). Rhyme generation in deaf students: the effect of exposure to Cued Speech. *J. Deaf Stud. Deaf Educ.* 8, 250–270. doi: 10.1093/deafed/eng014

Leybaert, J. (2000). Phonology acquired through the eyes and spelling in deaf children. *J. Exp. Child Psychol.* 75, 291–318. doi: 10.1006/jecp.1999.2539

Leybaert, J., and Colin, C. (2007). Le rôle des informations visuelles dans le développement du langage de l'enfant sourd muni d'un implant cochléaire. *Enfance* 59, 245–253. doi: 10.3917/enf.593.0245

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Laurence Erlbaum.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Nicholls, G. H., and Ling, D. (1982). Cued speech and the reception of spoken language. *J. Speech Hear. Res.* 25, 262–269.

Périer, O., Charlier, B., Hage, C., and Alegira, J. (1990). Evaluation of the effect of prolonged Cued Speech practice upon the reception of spoken language. *Cued Speech J.* IV, 45–59.

Rouger, J., Fraysse, B., Deguine, O., and Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Res.* 1188, 87–99. doi: 10.1016/j.brainres.2007.10.049

Schorr, E. A., Fox, N. A., Van Wassenhove, V., and Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18748–18750. doi: 10.1073/pnas.0508862102

Schwartz, J. L. (2010). A reanalysis of McGurk data suggest that audiovisual fusion on speech perception is subject-dependant. *J. Acoust. Soc. Am.* 127, 1584–1594. doi: 10.1121/1.3293001

Skinner, M. W., Fourakis, M. S., Holden, T. A., Holden, L. K., and Demorest, M. E. (1999). Identification of speech by cochlear implant recipients with the Multipeak (MPEAK) and Spectral Peak (SPEAK) speech coding strategies II. Consonants. *Ear Hear.* 20, 443–460. doi: 10.1097/00003446-199912000-00001

Sumby, W., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Troille, E. (2009). *De la Perception Audiovisuelle des Flux Oro-Faciaux en Parole à la Perception des Flux Manuo-Faciaux en Langue Française Parlée Complétée. Adultes et Enfants: Entendants, Aveugles ou Sourds*. Ph.D. dissertion, Université Stendhal-Grenoble III, Grenoble.

Troille, E., Cathiard, M. A., and Abry, C. (2007). *A Perceptual Desynchronization Study of Manual and Facial Information in French Cued Speech, ICPhS*. Saarbrücken.

Tyler, R. S., Parkinson, A. J., Woodworth, G. G., Lowder, M. W., and Gantz, G. J. (1977). Performance over time of adult patients using the Ineraid or nucleus cochlear implant. *J. Acoust. Soc. Am.* 120, 508–522.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 21 April 2014; published online: 19 May 2014.
*Citation: Bayard C, Colin C and Leybaert J (2014) How is the McGurk effect modulated by Cued Speech in deaf and hearing adults? Front. Psychol. **5**:416. doi: 10.3389/fpsyg.2014.00416*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

APPENDIX

Table A1 | TERMO scores by group and participant for Audio-Only, Visual-Only, AV, and Visual with CS (V + CS) cue conditions.

	Audio	Visual	AV	V + CS	Audio gain	CS gain
DEAF CS	60.38 (22.30)	35.88 (11.61)	77.00 (18.40)	94.75 (5.01)	41.13 (22.34)	58.89 (12.29)
1	82	24	82	94	58	70
2	12	29	41	100	12	71
3	71	35	94	88	59	53
4	59	35	88	100	53	65
5	65	47	94	88	47	41
6*	71	29	76	94	47	65
7*	47	59	59	100	0	41
8*	76	29	82	94	53	65
HEARING CS	100 (0.00)	38.14 (9.92)	100 (0.00)	88.43 (3.69)	61.86 (9.2)	50.29 (9.92)
1	100	47	100	88	53	41
2	100	41	100	88	59	47
3	100	41	100	88	59	47
4	100	24	100	82	76	58
5	100	29	100	88	71	59
6	100	41	100	82	59	41
7	100	35	100	88	65	53
8	100	59	100	94	41	35
9	100	18	100	94	82	76
10	100	35	100	88	65	53
11	100	41	100	94	59	53
12	100	41	100	88	59	47
13	100	41	100	88	59	47
14	100	41	100	88	59	47
HEARING CONTROL	99.60 (1.55)	36.27 (8.39)	100 (0.00)	42.33 (10.91)	63.73 (8.39)	6.07 (12.70)
1	100	41	100	47	59	6
2	100	47	100	47	53	0
3	100	41	100	47	59	6
4	100	29	100	35	71	6
5	100	47	100	59	53	12
6	100	35	100	35	65	0
7	100	18	100	47	82	29
8	100	29	100	53	71	24
9	100	35	100	41	65	6
10	100	41	100	53	59	12
11	100	29	100	53	71	24
12	100	29	100	24	71	−5
13	94	41	100	24	59	−17
14	100	35	100	29	65	−6
15	100	47	100	41	53	−6

*Standard deviations are indicated in parentheses. *Indicates participants with cochlear implants.*

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 August 2014
doi: 10.3389/fpsyg.2014.00934

Audiovisual spoken word training can promote or impede auditory-only perceptual learning: prelingually deafened adults with late-acquired cochlear implants versus normal hearing adults

Lynne E. Bernstein*, Silvio P. Eberhardt and Edward T. Auer Jr.

Communication Neuroscience Laboratory, Department of Speech and Hearing Science, George Washington University, Washington, DC, USA

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Mairead MacSweeney, University College London, UK
Riikka Mottonen, University of Oxford, UK

****Correspondence:***
Lynne E. Bernstein, Communication Neuroscience Laboratory, Department of Speech and Hearing Science, George Washington University, 550 Rome Hall, 810 22nd Street, NW Washington, DC 20052, USA
e-mail: lbernste@gwu.edu

Training with audiovisual (AV) speech has been shown to promote auditory perceptual learning of vocoded acoustic speech by adults with normal hearing. In Experiment 1, we investigated whether AV speech promotes auditory-only (AO) perceptual learning in prelingually deafened adults with late-acquired cochlear implants. Participants were assigned to learn associations between spoken disyllabic C(=consonant)V(=vowel)CVC non-sense words and non-sense pictures (*fribbles*), under AV and then AO (AV-AO; or counter-balanced AO then AV, AO-AV, during Periods 1 then 2) training conditions. After training on each list of paired-associates (PA), testing was carried out AO. Across all training, AO PA test scores improved (7.2 percentage points) as did identification of consonants in new untrained CVCVC stimuli (3.5 percentage points). However, there was evidence that AV training impeded immediate AO perceptual learning: During Period-1, training scores across AV and AO conditions were not different, but AO test scores were dramatically lower in the AV-trained participants. During Period-2 AO training, the AV-AO participants obtained significantly higher AO test scores, demonstrating their ability to learn the auditory speech. Across both orders of training, whenever training was AV, AO test scores were significantly lower than training scores. Experiment 2 repeated the procedures with vocoded speech and 43 normal-hearing adults. Following AV training, their AO test scores were as high as or higher than following AO training. Also, their CVCVC identification scores patterned differently than those of the cochlear implant users. In Experiment 1, initial consonants were most accurate, and in Experiment 2, medial consonants were most accurate. We suggest that our results are consistent with a multisensory reverse hierarchy theory, which predicts that, whenever possible, perceivers carry out perceptual tasks immediately based on the experience and biases they bring to the task. We point out that while AV training could be an impediment to immediate unisensory perceptual learning in cochlear implant patients, it was also associated with higher scores *during* training.

Keywords: cochlear implants, perceptual learning, multisensory processing, speech perception, plasticity training

INTRODUCTION

Pre-/perilingual severe or profound hearing impairment (henceforth, *deafness*) typically results in strong reliance on vision for communication, even in individuals who communicate with speech and regularly use hearing aids (Erber, 1975; Lamoré et al., 1998; Bernstein et al., 2000). Reliance on visual speech is observed also in individuals with cochlear implants (Giraud et al., 2001; Rouger et al., 2008; Huyse et al., 2013), particularly under noisy conditions that reduce the intelligibility of the auditory stimuli. The influence of vision in face-to-face communication or in audiovisual training with a cochlear implant could help in auditory perceptual learning, or it could hinder learning. This study was carried out to examine vision's influence during training that was intended to promote auditory perceptual learning.

Visual speech information could be beneficial to auditory perceptual learning if concordant visual speech information can guide the learning of new auditory input (Rouger et al., 2007). The use of another sense to guide learning to perceive the input from a sensory prosthesis is potentially a generalizable strategy. For example, sensory guided plasticity using auditory or vibrotactile stimuli has been suggested as a possible approach to enhancing perceptual learning with a visual prosthesis (Merabet et al., 2005). Audiovisual speech does provide concordant or correlated information (Jiang et al., 2002; Schroeder et al., 2008; Jiang and Bernstein, 2011) that is naturally available to the cochlear implant user. For example, easy visual distinctions such as "p" vs. "t," which are difficult auditory distinctions for the cochlear implant user, could be used to draw attention to potentially

available auditory distinctions and thereby promote learning. Evidence from studies of normal-hearing adults suggests that listeners are indeed able to use visual speech in learning novel auditory speech stimuli processed through a vocoder (Wayne and Johnsrude, 2012; Bernstein et al., 2013). But the retrospective evidence from studies on cochlear implant patients is mixed regarding the utility of audiovisual as opposed to auditory-only speech for auditory training (Bodmer et al., 2007; Dettman et al., 2013).

Neuroimaging evidence with normal-hearing adults and adults with cochlear implants suggests that individual differences as well as the quality of the visual and auditory input affect the extent to which auditory and visual speech input are processed (Nath and Beauchamp, 2012; Song et al., 2014): What an individual brings to a perceptual task, in combination with specific stimulus qualities, is important to the outcome of a perceptual task. The study reported here investigated how audiovisual training affects auditory-only speech perceptual learning in adults with prelingual deafness and late-acquired cochlear implants, and compared it with learning in adults with normal hearing.

LATE COCHLEAR IMPLANTATION

Cochlear implants are surgically implanted devices that deliver acoustic information to the cochlea to stimulate auditory neurons (Zeng et al., 2004). They use multiple channels of sound processing and multiple sites of stimulation along the length of the cochlea, mimicking to some extent the representation of frequencies by the normal cochlea (Wilson et al., 2011). Research on cochlear implantation in pre- and perilingually deafened children suggests that every year of delay in implantation during very early childhood is associated with reduced rates of language development (Niparko et al., 2010). From the earliest studies on cochlear implants there were consistent indications that late implantation is detrimental to outcomes, and that prelingually deafened children demonstrate an inverse relationship between age of cochlear implantation and magnitude of benefit from the implant (Waltzman et al., 1992; Snik et al., 1997; Manrique et al., 1999; Ponton et al., 1999; Sharma et al., 2002; Teoh et al., 2004). Such results have been interpreted as evidence for a critical period for successful cochlear implantation of children (Snik et al., 1997; Sharma et al., 2002), beyond which plasticity is closed (Ponton et al., 1996; Fryauf-Bertschy et al., 1997; Knudsen, 2004; Kral and Sharma, 2012).

However, particularly in the past decade, opinion seems to have shifted toward support for the possibility that there is benefit associated with late cochlear implantation (Osberger et al., 1998; Waltzman and Cohen, 1999; Teoh et al., 2004; Moody-Antonio et al., 2005). Teoh et al. (2004) conducted a retrospective study of 103 adult patients in clinical trials and a meta-analysis of all published studies of patients with pre-lingual deafness and cochlear implants. Patients had onset of deafness at less than 3 years of age and cochlear implantation at greater than 13 years of age. In the first year, mean auditory-only performance on sentences in quiet was approximately 30% words correct, Hearing in Noise Test (HINT) (Nilsson et al., 1994) sentences in quiet were approximately 20% words correct, and monosyllabic words in quiet were approximately 15% words correct. Individual scores on the HINT test ranged between 40 and 100% correct for a subset of individuals. No significant differences were found among implant hardware or processors, leading the authors to conclude that "patient characteristics, rather than device properties *per se*, are likely to be the major contributing factor responsible for the outcome measures" (p. 1539).

Waltzman et al. (2002) reported on 14 congenitally deaf adults (with mean age 26 years). Scores on speech measures varied widely. For example, pre-operatively auditory-only scores on monosyllabic words were in the range 0 to 12% correct and post-operatively were in the range 0 to 46%. Pre-operatively, scores on sentences were in the range 0 to 38% words correct in quiet and post-operatively were in the range 0 to 98% correct. Pre-implant performance did not predict post-implant scores. Residual hearing was rejected as a predictor for favorable outcomes, but newer processing algorithms along with reliance on oral speech and language were considered to be potentially important. Schramm et al. (2002) also reported benefits and wide individual differences. Fifteen patients, implanted across the age range 12–49 years, exhibited scores on isolated auditory-only sentences post-implant from 0 to 98% correct. Suggested factors for individual differences included the age at time of implant, extent of therapy, overall experience in an oral environment, patient/family motivation and support systems, degree of residual hearing before implantation, and level of auditory functioning before implantation.

In Moody-Antonio et al. (2005), we reported on auditory-only, visual-only, and audiovisual scores for words in unrelated sentences presented to eight prelingually deafened adults with late-acquired cochlear implants. Even with essentially no auditory-only speech perception, some individuals were able to show enormous audiovisual gains over their visual-only scores. Similarly, in a recent study (Bodmer et al., 2007) that included 109 English-speaking adult cochlear implant patients who were pre-/perilingually deafened, 24 were placed in the category of excellent implant users. They had all received strong auditory or oral education that was said to include use of visual speech.

Thus, the emerging picture suggests that pre-/perilingually deaf adults with speech communication experience can benefit from a cochlear implant, even if it is obtained after what might be considered a critical period for first language acquisition and speech perception. There is evidence that even if their auditory-only speech perception is poor, some pre-/perilingually deafened individuals can benefit from a cochlear implant by combining auditory and visual information. However, the ability to combine visual and auditory speech features to carry out a perceptual task is not identical to using visual perception to improve auditory speech perception. To our knowledge, the question of whether visual information promotes auditory perceptual learning has not heretofore been studied experimentally with this clinical population.

THIS STUDY

The design of this study used elements of the training experiments reported in Bernstein et al. (2013). Training was given in a paired-associates paradigm for which the task was to learn to associate spoken CVCVC (C = consonant, V = vowel) non-sense

words with so-called *fribble* non-sense object pictures (Williams and Simons, 2000). The modality during paired-associates training was either audiovisual (AV) or auditory-only (AO), but testing on the paired-associates was always AO, and it always followed immediately after training on each list of paired-associates. This task requires establishing semantic relationships between spoken words and pictures, and learning the auditory stimuli well enough to demonstrate knowledge of the semantic relationship when the spoken words are AO, regardless of whether they were trained with AO or AV stimuli.

Participants were assigned to two different orders of training (i.e., referred to as "modality assignments"), AV-AO with AV first, or AO-AV with AO first. Their task during each modality assignment was to learn three lists of 12 paired-associates. Prior to training, between the switch to a different training modality, and following both modalities of training, they identified consonants in untrained sets of AO CVCVC non-sense words. Experiment 1 applied these methods with prelingually deafened individuals with late-acquired cochlear implants, and Experiment 2 applied the same methods to normal-hearing adults.

MATERIALS AND METHODS

EXPERIMENT 1: COCHLEAR IMPLANT PARTICIPANTS

Cochlear implant participants

Individuals were recruited through the House Clinic (Los Angeles, CA). Individuals were screened for American English as a first language and normal or corrected-to-normal vision in each eye of 20/30 or better (using a Snellen chart). The recruitment goals were pre- or perilingual profound hearing loss and a late cochlear implantation. Late implantation was considered to be 5 years of age or older. The total number of initially enrolled cochlear implant patients was 33. Twenty-eight are included in this report. Of the 5 excluded the reasons for exclusion were: One participant received incorrectly ordered training blocks, two discontinued the study after the first day, and two were identified as deaf at age 5 years. The included participants ranged in age from 20 to 53 years (mean = 37.1 years), with 15 males.

Table 1 shows that most participants were diagnosed as deaf at birth, mostly of unknown origin, although records showed that the hearing loss onset was 3 years of age for one participant but was deemed likely progressive from birth. Cochlear implant activation age was 6 or 8 years for three of the participants. Most of the implants were obtained beyond 19 years of age. Implantation as young as 6 years is not considered problematic for this study, because implantation after even the second or third birthday is associated with far worse outcomes than for younger patients, and the odds for good results with an implant are considered to be very poor after 4–6 years of age (Kral and Eggermont, 2007; Wilson et al., 2011). That three participants used bilateral cochlear implants was not considered problematic in light of evidence that the additional implant may be of marginal benefit (Yoon et al., 2011), and there is no reason to believe that the task would benefit from two rather than one implant.

All of the participants had hearing aid experience at some time in their lives. But pure tone average scores were obtained using only their implant, and only cochlear implants were used during the study. **Table 1** lists the type of implant used during the experiment.

Participants were tested with the Peabody Picture Vocabulary Test (PPVT) (Dunn and Dunn, 1981) and the Comprehensive Test of Non-verbal Intelligence (C-TONI) (Hammill et al., 1996). All participants received a lipreading screening test (Auer and Bernstein, 2007).

Participants were paid $12 per hour plus any travel expenses incurred. The entire experiment was generally carried out across 2 days of testing at the House Research Institute (Los Angeles, CA). Participants gave written consent. Human subject participation was approved by the St. Vincent's Hospital Institutional Review Board (Los Angeles, CA).

STIMULI

All visual and auditory stimulus materials were identical to those used in Bernstein et al. (2013). All of the words and word lists are presented in that publication. A brief description of the stimuli is provided here for convenience.

Speech

The spoken CVCVC (C = consonant, V = vowel) non-sense words used for the paired-associates training and testing, as well as for the consonant identification task, were modeled on English phonotactics (i.e., the sequential speech patterns in English) using Monte Carlo methods. There were 260 unique words, which were recorded with a female talker. All of the words were visually distinct for lipreading and also visually unique from real English words (i.e., the words were designed to not be mistaken as real words, if they were lipread without accompanying audio). Thus, for example, the non-sense word *mucker* was not included in the set, because the visual stimulus could be mistaken for the real word *pucker*, inasmuch as the phonemes /p, m/ are visually highly similar (Auer and Bernstein, 1997). The full set of non-sense words includes all the English phonemes, and within each CVCVC, the five phonemes are expected to be visually distinct to a lipreader. Recently obtained results (Eberhardt et al., submitted) show that the stimuli can be learned in the paired-associates paradigm described below using only the video stimuli. Two 49-item lists were selected for the consonant identification task (see below). Two six-item lists were selected for pre- and post-training practice. Six lists of 12 items for paired-associates training and six lists of 6 items as new items during PA testing were selected from the remaining available words. The three stimulus lists for AV training were the same three lists regardless of when AV training was given, and the same was done for the three AO training lists. In other words, order of training was counter-balanced, but list was locked with only the AO or AV training modality. No evidence of list effects (in terms of items within lists) was observed previously (Bernstein et al., 2013).

Non-sense pictures

Non-sense pictures in the PA task were from the fribbles image set (http://wiki.cnbc.cmu.edu/Novel_Objects). Fribbles comprise 12 species with distinct body "core" shape and color, with 81 exemplars per specie obtained by varying the forms of each of four appendage parts. From the available images, six lists of 12 images

Table 1 | Experiment 1 participants.

Participant	Modality assignment	Age of onset (years)	Etiology	Age at activation in years	Age at testing in years	Pure tone Ave. (dBHL)	Pretest initial consonants correct (%)	Implant	Lipreading words correct (%)
1	AO-AV	Birth	Rubella	33	44	33	41	CI24M +	45
2	AV-AO	1.1	Unknown	47	48	30	67	Clarion II	56
3	AV-AO	3.0	Unknown	19	22	13	16	CI24M	63
4	AO-AV	0.5	Genetic	6	20	27	31	CI22M	29
5	AO-AV	1.3	Meningitis	46	53	23	41	CI24R	59
6	AV-AO	0.5	Unknown	43	44	27	37	CI24RE	42
7	AO-AV	1.5	Rubella	39	45	30	31	CI24R+	56
8	AV-AO	1.5	Rubella	46	47	15	82	CI24RE	53
9	AO-AV	2.0	Rubella	52	52	15	69	CI512	63
10	AO-AV	0.3	Meningitis	51	51	20	12	CI512	37
11	AO-AV	Birth	Unknown	43	44	20	59	CI24RE	14
12	AO-AV	1.0	Mondini	6	23	31	37	CI22M	50
13	AV-AO	Birth	Genetic	26	34	37	43	Clarion II	27
14	AV-AO	0.8	Rubella	30	35	20	18	CI24M	65
15	AO-AV	2.0	Meningitis	23	24	17	16	CI512	16
16	AV-AO	Birth	Prematurity	45/48	52	23	06	CI24R/RE+	13
17	AV-AO	Birth	Unknown	30	35	24	04	Clarion 90K	24
18	AO-AV	Birth	Rubella	35	35	32	57	CI512	49
19	AO-AV	Birth	Genetic	17	25	15	08	CI24M	41
20	AO-AV	1.0	Fever	16	22	40	57	CI24R	53
21	AO-AV	Birth	Unknown	7	20	40	47	CI24M	33
22	AO-AV	Birth	Unknown	34	43	25	14	Clarion II	36
23	AV-AO	Birth	Unknown	19	26	20	29	CI24M	26
24	AV-AO	Birth	Unknown	23	29	15	10	CI24M	45
25	AV-AO	Birth	Unknown	42	52	21	57	Clarion II	56
26	AO-AV	Birth	Unknown	13	23	37	06	Clarion S	59
27	AV-AO	Birth	Genetic	38	49	33	22	Clarion II	53
28	AO-AV	Birth	Unknown	34	41	32	31	Clarion II	31
				30(14)†	37(12)	26(8)	34(22)		43(16)

*Notes: †Mean (standard deviation); +bilateral implants; *unknown/multiple strategies.*

The table lists each participant's modality assignment for paired-associates training and test, their age of deafness onset in years, its etiology, the age at which their cochlear implant was activated, the age when tested, their pure tone average with the cochlear implant, their score for the initial phoneme in CVCVC stimuli at the pre-test, the type of implant, and their lipreading screening score.

each were created such that each list used three different body forms and no duplicated appendage forms, rendering the images within each list highly distinctive (Williams and Simons, 2000). No appendage was repeated across lists.

OVERALL DESIGN OF THE PROCEDURE

Figure 1 shows the overall design of the experiment. Participants were assigned to either AO-AV (i.e., AO first, AV second) or AV-AO orders of paired-associates (PA) training. Within each assignment, the first list was trained for three blocks (giving six pseudorandom presentation of each of the twelve CVCVC words), and the next two lists for two blocks only[1]. Each list was tested AO immediately after training. Participants also carried out consonant identification with CVCVC non-sense words on three occasions, before training (pre-1), after the first set of three lists (post-1), and after the second set of three lists (post-2).

Paired-associates training procedure

Figure 2, repeated from Bernstein et al. (2013), outlines the events within a PA training trial. During training, the participant's task was to learn, by trial and error with feedback on each trial, lists of individual associations between each of 12 CVCVC spoken non-sense words and 12 fribble images. In the figure, an AV training

[1]The decision to train for three blocks on the first list and then two blocks for each of the subsequent lists within a modality assignment was because the participants were available for only two separate days; and the goal was to train on several lists within the available time period. The three-block training for the first list was considered important for establishing task learning. In addition, previous training with normal-hearing participants suggested the possibility that three blocks of training could produce ceiling performance, thereby reducing our ability to detect modality effects. However, the results of Experiment 1 showed this not to be the case for the cochlear implant users.

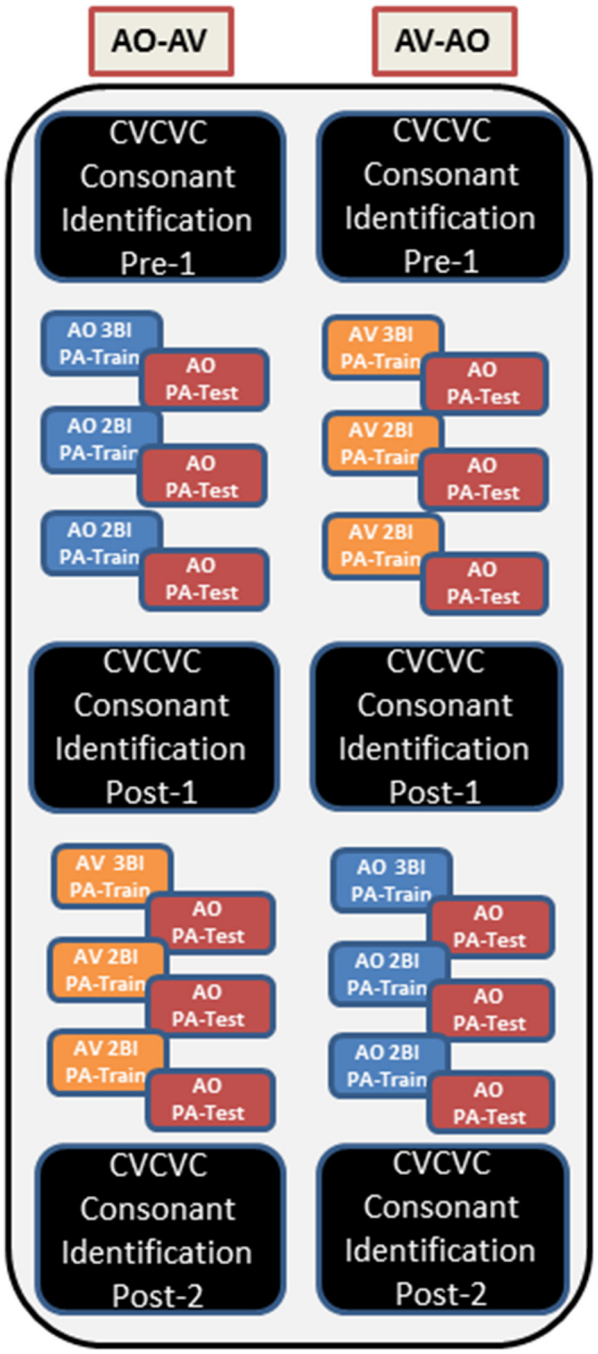


FIGURE 1 | Overall design of Experiments 1 and 2. The left column corresponds to the AO-AV modality assignment, and the right column corresponds to the AV-AO modality assignment. Within each assignment there were six training lists, three with AV stimuli and three with AO. The first list in each modality was trained for three blocks (BI), and the subsequent two lists were trained for two blocks. At pre-1, post-1, and post-2 times, CVCVC consonant identification was tested using untrained stimuli.

trial is shown in the left column and an AO training trial is shown in the right column. Each trial began with a computer-monitor display of the 12-fribble image matrix (3 rows of 4 columns, with image position within the matrix randomly selected on a trial-by-trial basis). During AV training, a video of the talker was played in synchrony with the spoken audio, and during AO training, a single still image of the talker's face was displayed on the monitor during audio presentation. The talker was presented on a different monitor than the fribble matrix monitor, and a large arrow appeared on the bottom of the fribble monitor pointing left to remind the participant to focus attention on the talker. The participant used the computer mouse to choose a fribble image following the speech stimulus. Feedback was given by outlining the correct fribble in green and an incorrect choice in red. After a short interval, the speech stimulus was always repeated, while the fribble images and borders remained unchanged.

A training block comprised two (or three for List 1) repetitions of the twelve paired associations in pseudorandom order. Prior to the first training list in each condition (AV or AO), participants were given practice with one block of 6 trials. The training score was the proportion of correct paired associations of trained words in the block.

Paired-associate testing procedure

Paired-associates testing immediately followed training. The testing procedure was the same as that of training, except the CVCVC speech stimuli were always presented AO, no feedback was given, the stimulus was not repeated during the trial, and each response triggered the next trial. Six of the trained words and all 12 of the fribble images were used for testing. The associations for the six retained words were unchanged. Six foil CVCVC non-sense words were paired with the fribble images of the discarded words. A testing block comprised, in pseudorandom order, four presentations of the twelve stimuli. The test score was the proportion of correct paired associations of the six originally-trained words across all trials.

CVCVC phoneme identification

In a forced choice paradigm, participants identified the three consonants in 49 different CVCVC stimuli before their first training period (pre-1), after their first training period (post-1), and after their second training period (post-2). The CVCVC stimuli had varied vowels that were not identified and 24 possible consonants transcribed using the computer keyboard and single characters from ARPABET, /b, d, f, g, h, k, l, m, n, p, r, s, t, v, w, y, z, C, D, G, J, S, T, Z/ (which correspond to the International Phonetic Alphabet, /b, d, f, g, h, k, l, m, n, p, r, s, t, v, w, j, z, tʃ, ð, ŋ, dʒ, ʃ, ɵ, ʒ/). These CVCVC stimuli were all different from those in the paired-associates training paradigm.

In order to familiarize participants with the transcription set, they were given a chart that showed each of the ARPABET symbols, and they filled out two worksheets with words spelled using English orthography. Each word had a consonant underlined, and the participant transcribed the underlined letter using the correct ARPABET symbol. The first worksheet was filled out with access to the chart and the second without. Mistakes were corrected, and other examples were given by the research assistant who worked with participants until the participant was comfortable using the symbol set. During testing the participants could see a chart with the ARPABET symbols and word examples on the computer screen. The three consonant positions were marked

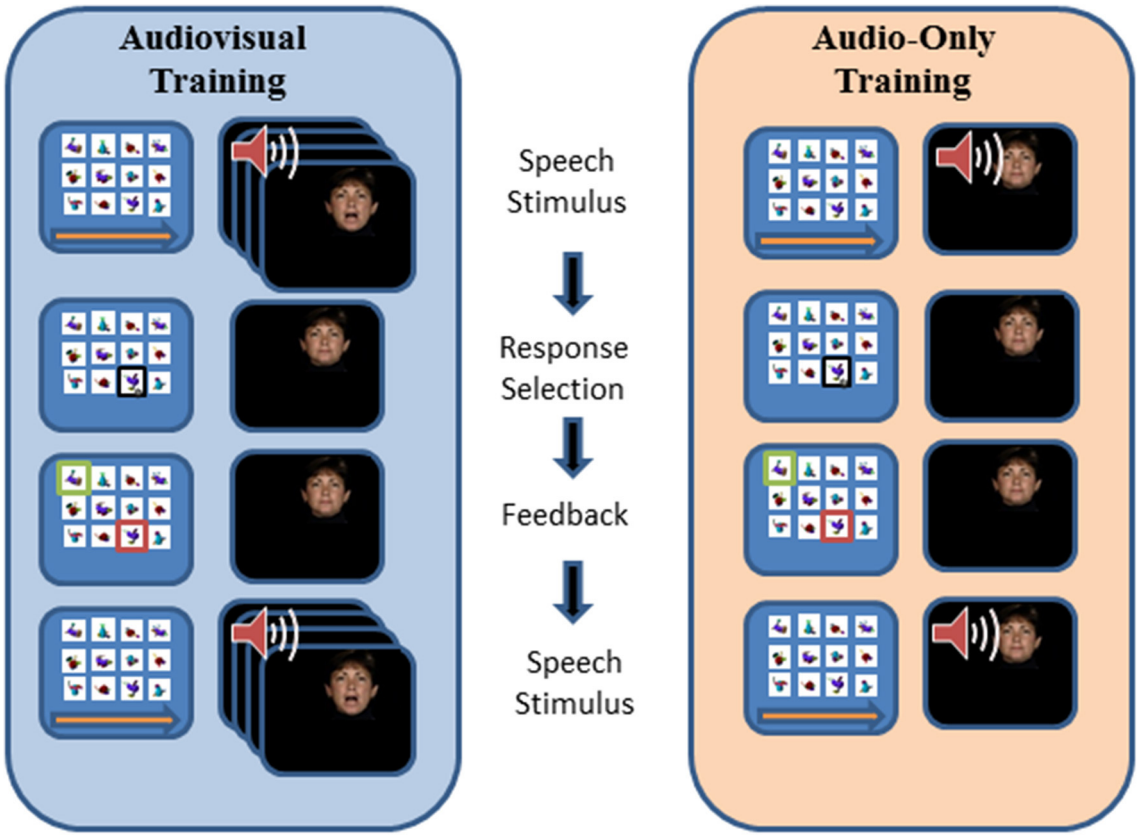


FIGURE 2 | Sequence of events during a paired-associates training trial with AV training (on the left) and AO (on the right) stimuli. A speech stimulus was presented, followed by the participant's response selection, followed by feedback and a repetition of the speech stimulus. Each panel depicts the screen showing the fribble images side-by-side with the video monitor showing the talker. The trial structure for AV and AO training followed the same sequence, except that during AV training the video was played synchronously with the audio, and during AO training a still neutral face was played during the audio (adapted from Bernstein et al., 2013).

on the computer screen with "__-__-__" and the participants used the keyboard to fill in the blanks. They could backspace and correct mistakes. They were given a practice list prior to starting each test list. There were two unique lists of CVCVC stimuli, A and B, and these lists were counter-balanced across participants so that they received either ABA or BAB list orders across the pre-1, post-1, and post-2 tests. The task resulted in a percent correct score for each consonant position in the CVCVC stimuli.

APPARATUS

Audiovisual CVCVC tokens were digitized, edited, and conveyed to digital video disk (DVD) format. The participants listened in the sound field. The audio stimuli were output at a calibrated 65 dB A-weighted sound pressure level (SPL) using a JBL LSR6325P-1 loudspeaker. Cochlear implant thresholds were checked using audiometry prior to participating in each test session. Testing took place in an IAC (Industrial Acoustics Company) double-walled sound-attenuating booth using a standard computer interface that included a 51 cm LCD monitor, and a 35.6 cm Sony PVM-14N5U NTSC video monitor for display of speech video from the DVD. Monitors were located about 1 m from the participant's eyes, so that the computer monitor subtended a visual angle of 23.1° horizontally and 17.3° vertically with the 12 fribble matrix filling the monitor. The visual speech was displayed on the NTSC monitor with the talker's head subtending visual angles of 3.9° horizontally and 5.7° vertically. Custom software was used to run the experiment, collect responses, and compile data.

ANALYSES

All responses were converted into proportions correct and then arcsine transformed, $y = \sin^{-1}(\sqrt{p})$, where p is the proportion correct. This transformation addressed the analyses of variance sphericity requirement given proportion scores across the range 0 to 1.0. Untransformed data failed to pass Mauchley's test of sphericity, and thus the variance differences of untransformed data were different. The score range following the arcsine transformation is 0 to 90. Statistics are reported on the arcsine transformed data, but tables, means, and figures present untransformed data to facilitate interpretation. Multivariate analyses of variance, simple contrasts, and t-tests were carried out with SPSS (IBM Statistics SPSS 22). Unless explicitly noted, only effects that were reliable at the level of $p < 0.05$ are reported.

RESULTS

Participant characteristics

Independent samples *t*-tests were used to determine whether there were participant characteristic differences between the AO-AV and AV-AO modality assignment groups (see **Table 1**). There were no differences found between groups in terms of scores on lipreading screening, PPVT scores, TONI scores, duration of time between acquiring the cochlear implant and participation in the study, age of cochlear implant activation, age of hearing loss onset, initial consonant percent correct in CVCVC stimuli pre-training, age at testing, or pure tone average (each, $p > 0.085$). The initial consonant scores in the pre-training CVCVC consonant identification test was compared across groups to probe whether auditory speech perception differed across groups prior to training, and it did not. The initial consonant was used because it was deemed a reasonable check on pre-training auditory speech perception.

Potential covariates with paired-associates training and test scores

While the analyses of participant characteristics showed that the AO-AV and AV-AO groups were not different in terms of the various individual participant characteristics, scores could vary systematically with the training or test measures. Bivariate correlations were tested between individual participant characteristics (i.e., lipreading screening scores, PPVT scores, duration of time between acquiring the cochlear implant and participation in the study, age of cochlear implant activation, age at testing, and pure tone average) and the 10 training and test scores for each type of modality assignment (i.e., 10 scores for the AV training and testing, and 10 for the AO training and testing). None of these individual participant characteristics was reliably correlated with the training or test scores. Therefore, none of the individual participant characteristics was used as a covariate in any of the foregoing statistical analyses.

Overview of the paired-associates training and testing time series

Figure 3A shows the time series of mean training and test scores for every training block and test block across the two modality assignments (AV-AO, AO-AV) in Experiment 1. The figure suggests that training scores during Period 1 were similar for AV and AO assignments, but AO test scores were lower following AV training. In contrast, the Period-1 AO training block scores preceding test were similar to the test scores on the same list. In addition, the figure suggests that the times series across the two periods of training and testing varied depending on the order of AV vs. AO training assignment, with AO-AV participants turning in the better training performance during Period 2. The figure also suggests that the AO test scores following AV training were reduced relative to training scores. In light of this apparently

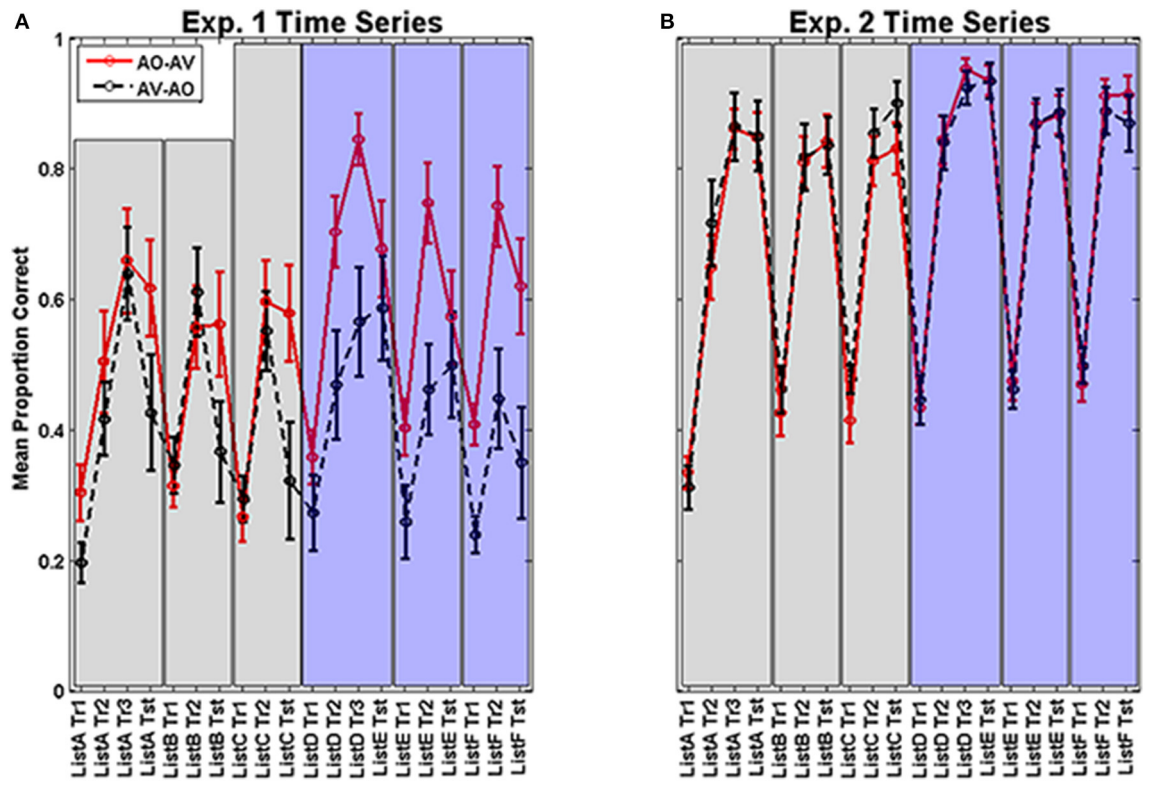


FIGURE 3 | (A,B) Experiment 1 (left **A**) and 2 (right **B**) time series. Means and standard errors of the mean for each training block and test block are shown across Period 1 (left half of each figure) and Period 2 (right half of each figure) for the AV-AO and AO-AV modality assignments. Lists were in a fixed assignment within each modality, so list designations in the figure (*x*-axis) represent the points in time for each training and test block. (Note: Tr, Training Block 1; and Tst, Test).

complex pattern of results, statistical analyses were carried out first on the training scores, then the test scores, and then the difference scores that were calculated between the final training block and its subsequent test block for each of the six training lists.

Paired-associates training scores results

Analyses of the training results were carried out using the last training block per list, because the final training block gives an estimate of best performance in the training condition. Analysis was carried out with the within-subjects factors of training list (3) and training period (Period 1: first three assigned lists; Period 2: second 3 lists), and the between subjects factor modality assignment (AO-AV: AO in Period 1 followed by AV in Period 2; and *vice versa*, AV-AO). MANOVA showed that list was a reliable main effect, $F_{(2,\ 25)} = 5.705$, $p = 0.009$, $\eta_p^2 = 0.313$, independent of modality assignment. List scores dropped reliably from List 1 (mean = 67.7%) to 2 (mean = 59.4% correct), $F_{(1,\ 26)} = 8.437$, $p = 0.007$, $\eta_p^2 = 0.245$, but not from List 2 to 3 (mean = 58.5% correct) ($p = 0.891$). List did not interact with any other factors.

The main effects of training period and modality assignment were not statistically significant. However, training period and modality assignment interacted, $F_{(1,\ 26)} = 19.711$, $p = 0.000$, $\eta_p^2 = 0.431$, as was suggested by the time series shown in **Figure 3A**. This interaction is shown in **Figure 4**, for which the pooled means that entered into the interaction are graphed. There was an improvement for AO-AV participants' training scores between AO and AV training periods vs. a decline in AV-AO participants' training scores. The Period-1 training scores were mean 60.2% correct across both modality assignments. The Period-2 scores for the AO-AV modality assignment were 77.8% correct and for the AV-AO modality assignment, 49.2% correct; that is, there was a 28.6 percentage point difference between the AV vs. AO training scores during Period 2.

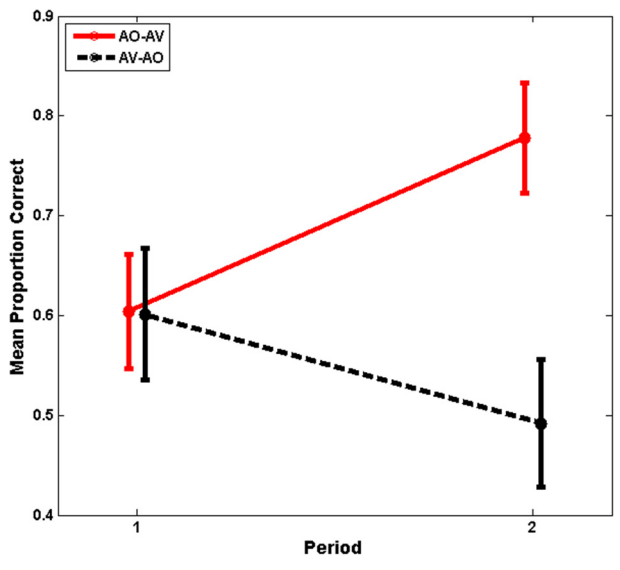


FIGURE 4 | Experiment-1 mean training scores with standard errors of the mean. Period-1 training scores were the same independent of training modality, but training scores diverged at Period 2, with higher mean scores during AV training.

The interaction between training period and modality assignment was then investigated. Of interest was whether the increase in training scores across training Periods 1 and 2 for the AO-AV modality assignment and the decrease in training scores for the AV-AO assignment were both reliably different from zero. Indeed, the increase across periods for the AO-AV assignment was reliably different from zero, $p = 0.000$; but the decrease for the AV-AO assignment was a marginal drop, $p = 0.062$. For completeness, it is noted that the change between Period 1 to Period 2 also differed across modality assignments: AO-AV scores increased 17.4 percentage points, and AV-AO scores declined 10.9 percentage points, differing across groups, $t_{(26)} = 4.440$, $p = 0.000$. Thus, during Period 1, training resulted in similar performance regardless of training modality; but the training scores rose significantly across period for the AO-AV group and fell marginally for the AV-AO group.

Paired-associates test results

Following training on each list, participants were tested AO on the number of paired-associates they had learned. Testing used 6 out of the 12 trained associations and 6 untrained foils. The test scores for the trained words were submitted to an omnibus analysis with the within-subjects factors test list (3) and testing period (Period 1, first 3 lists; Period 2, second 3 lists), and the between-subjects factor modality assignment (AO-AV, AV-AO).

The main effect of testing period was reliable, $F_{(1,\ 26)} = 5.500$, $p = 0.027$, $\eta_p^2 = 0.175$. Period 1 test scores were mean 47.9% correct. Period 2 scores were mean 55.1% correct. Overall, the mean difference across periods was 7.2 percentage points.

List scores also differed, $F_{(2,\ 25)} = 6.805$, $p = 0.004$, $\eta_p^2 = 0.352$; and simple contrasts showed that independent of training modality, scores declined from List 1 to 2, $F_{(1,\ 26)} = 8.632$, $p = 0.007$, $\eta_p^2 = 0.249$, and List 2 and 3 scores were similar (List 1, 57.7%; List 2, 50.1%; List 3, 46.8%). This effect was not surprising as it mirrored the list effect obtained with training scores.

But the modality assignment by list interaction was also reliable, $F_{(2,\ 25)} = 3.520$, $p = 0.045$, $\eta_p^2 = 0.220$. AV-AO test scores dropped between Lists 2 and 3 relative to those of the AO-AV participants, $F_{(1,\ 26)} = 7.299$, $p = 0.012$, $\eta_p^2 = 0.219$. AO-AV participants' scores increased from List 2, 56.8% to List 3, 59.9%, but AV-AO participants' scores dropped from 43.3 to 33.7% correct across Lists 2 to 3.

The individual time series test scores in **Figure 3A** suggest that modality assignment had a differential effect on AO tests scores during Period 1 (**Figure 5** shows the mean test scores with standard errors of the mean.). As we do later in Experiment 2, we considered the Period-1 scores to be the best estimates of how AV vs. AO training affects AO learning; because at Period 2, the participants' training is conditioned on different experiences within the study. As a consequence, Period 2 cannot be used to estimate training modality *per se*. An analysis was carried out on the Period-1 scores, with the within-subjects factor list (3) and the between-subjects factor modality assignment. In that analysis, list and list by condition were not reliable effects. The condition effect returned the statistics, $F_{(1,\ 26)} = 4.175$, $p = 0.051$, $\eta_p^2 = 0.138$.

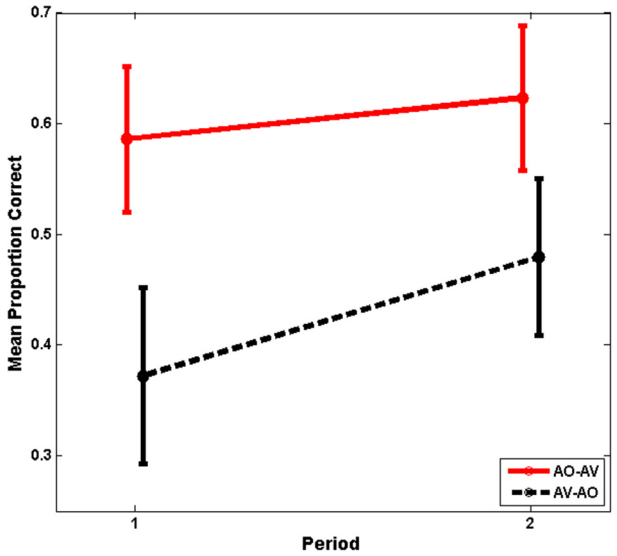


FIGURE 5 | Experiment-1 mean test scores with standard errors of the mean. Period-1 test scores were lower for AV-trained participants, whose scores improved significantly in Period 2.

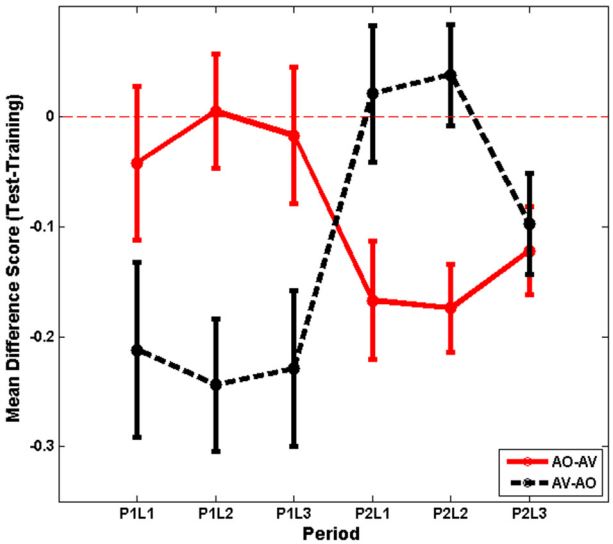


FIGURE 6 | Experiment 1 time series for mean difference scores (test minus training score) with standard errors of the mean per list shown separately for AO-AV vs. AV-AO modality assignments (P1L1, Period 1, List 1).

In this case, the result without application of the arcsine transform was more reliable, $F_{(1,\ 26)} = 4.367$, $p = 0.047$, $\eta_p^2 = 0.144$. Neither analysis violates Mauchley's test of sphericity. But the more conservative approach is associated with a slightly elevated possibility that we may incorrectly reject the null hypothesis. The mean AO test scores for AV-trained participants in Period 1 was 37.2% correct. The mean AO test scores for AO-trained participants in Period 1 was 58.6% correct.

Paired-associates training vs. test scores compared

A different approach to evaluating training is to consider the relationship between training and test scores. Across training and test results, there was a pattern of greater stability between AO training and test scores than between AV training and AO test scores (see **Figure 3A**). To investigate this pattern, the test minus training difference scores were calculated per participant for each list (6). **Figure 6** shows the time series of the difference scores separated across modality assignment group (AO-AV, AV-AO).

Differences scores were submitted to analysis with the within-subjects factors list (3) and training period (2) and the between-subjects factors training assignment (AO-AV, AV-AO). The interaction between training period and modality assignment was the only reliable effect, $F_{(1,\ 26)} = 16.295$, $p = 0.000$, $\eta^2 = 0.385$. Participants in the AO-AV assignment dropped 1.8 percentage points between training and test during their AO assignment; and then during their AV assignment their scores dropped 15.5 percentage points going from training to test. Participants in the AV-AO assignment dropped 22.8 percentage points between training and test during their AV assignment; and then during their AO assignment their scores dropped 1.3 percentage points between training and test.

The analysis above of training scores had shown that the Period-1 training scores did not vary between the AV- and AO-trained groups. A question then was whether the declines experienced with AV training varied across period, and they did not, $p = 0.341$. Thus, there was no evidence that either order of AV training resulted in a less steep decline from AV training to AO testing.

CVCVC forced-choice consonant identification

In a forced choice paradigm, participants identified the three consonants in CVCVC stimuli before their first training period (pre-1), after their first training period (post-1), and after their second training period (post-2). Their proportion correct scores were computed separately for each consonant position (initial, medial, final) and each period of testing (pre-1, post-1, and post-2). These scores were submitted to analyses for within subjects factors CVCVC testing period (3), and consonant position (3), and for the between subjects factor training modality assignment (AV-AO, AO-AV). **Table 2** shows the consonant identification mean scores for each period of testing and modality assignment.

The two main effects of test period, $F_{(2,\ 25)} = 8.015$, $p = 0.002$, $\eta_p^2 = 0.391$, and position, $F_{(2,\ 25)} = 6.876$, $p = 0.004$, $\eta_p^2 = 0.355$, were reliable. Simple comparisons showed that post-2 scores (32.3% correct) were higher than post-1 scores (29.6%), $F_{(1,\ 26)} = 5.816$, $p = 0.023$, $\eta_p^2 = 0.183$. Between pre-1 (28.9%) and post-2 the scores improved overall 3.4 percentage points. Simple comparisons also showed that consonant Position 1 scores (33.6% correct) were higher than Position 2 scores (29.2% correct), $F_{(1,\ 26)} = 13.913$, $p = 0.001$, $\eta_p^2 = 0.349$. But there was not a difference between Positions 2 and 3 (28.0% correct). Between Positions 1 and 3 the difference in scores was 5.6 percentage points.

Discussion

In Experiment 1, two participant groups, who did not differ in terms of various individual measures such as lipreading screening

Table 2 | Phoneme identification scores in Experiments 1 and 2.

Experiment	Pre-1			Post-1			Post-2		
	Initial	Medial	Final	Initial	Medial	Final	Initial	Medial	Final
1—Cochlear implant users	0.328 (0.043)	0.280 (0.038)	0.260 (0.040)	0.337 (0.045)	0.286 (0.036)	0.266 (0.038)	0.345 (0.047)	0.310 (0.040)	0.313 (0.040)
2—Normal-hearing	0.296 (0.009)	0.458 (0.020)	0.319 (0.014)	0.390 (0.012)	0.587 (0.018)	0.429 (0.014)	0.419 (0.011)	0.640 (0.018)	0.472 (0.016)

The table gives the mean (standard error of the mean in parentheses) for initial, medial, and final consonants correct scores for pre-1, post-1, and post-2 testing times for the prelingually deaf late-implanted cochlear implant users (Experiment 1) and the normal-hearing adults (Experiment 2).

scores and duration of cochlear implant use, were trained using AV and AO stimuli in a design for which the order of AV or AO training was counter-balanced across groups. But all testing was carried out with AO stimuli.

Period 1 was the better one to estimate the effect of training modality, because scores were not conditioned on prior training experience in the experiment, as was the case during Period 2. During Period 1, training scores were similar across groups, regardless of whether their training was AV or AO. However, AV-trained participants' AO test scores were lower than their training scores by an average 22.8 percentage points; while the AO-trained participants' AO test scores stayed essentially the same at test (1.8 percentage points different between training and testing). Given similar training scores across groups during Period 1, the lower AO test scores following AV training do not seem attributable to poorer ability for learning paired associations. In fact, a *post-hoc* paired-samples *t*-test shows that the AV-AO participants were capable of much better AO test performance when it followed AO training, $t_{(11)} = 2.570$, $p = 0.026$ (Period-1 mean, 37% correct; Period-2 mean, 48% correct).

Although the difference scores between training and test are better indicators of the effect of training on individual participant's performance, we also evaluated the AO test scores. While the results based on arcsine transformed scores are associated with a slightly elevated risk of falsely rejecting the null hypothesis ($p = 0.051$, raising the risk by 0.001) whereas the analysis based on untransformed scores was reliable (at $p = 0.047$), the AO test score analysis also showed that AV training is worse than AO training for learning the AO stimuli.

During *Period 2*, again a large drop between AV training and AO test scores (11.5 percentage points) was observed. The Period-2 pattern of results is, however, less amenable to straightforward interpretation, because the modality of the previous Period-1 training experience necessarily influenced performance. In Period 2, the participants brought different experience to the training and test tasks. For example, learning the training task with AO stimuli in Period 1 may have helped to focus on the auditory part of the AV stimuli during training in Period 2.

The drop between AV training and AO testing is complicated to interpret, in part because we do not have an independent estimate of what might be the "most accurate" AO test performance that could be achieved with a cochlear implant. The drop in scores from AV training to AO testing during Period 2 might for example be due to transducer limitations intrinsic to the cochlear implant. If so, Period-2 AO-AV test scores might have approached the best performance possible without also being able to see the talker. Repeated training and AO testing would be needed to obtain asymptotic performance for estimating the magnitude of the contribution afforded by visible speech beyond the available auditory information. Furthermore, to pin down the roles of amount of training, order of training, and modality of training, a control experiment is needed that includes AO-AO and AV-AV training in an expanded design with a new set of cochlear implant users of the type here and random assignment to groups.

In Experiment 1, participants also were tested on identification of consonants in untrained CVCVC non-sense words, and overall scores improved 3.4 percentage points. This result suggests that generalization took place beyond the word-learning task. If participants had merely learned non-sense words as holistic units, they should not have improved their scores on the untrained CVCVC stimuli. In addition, the consonant identification test scores suggest a bias based on visual speech perception, which is discussed in the General Discussion section. This bias is perhaps related to lipreaders' more accurate perception of initial position consonants in CVCVC stimuli (Auer and Bernstein, in preparation). On the other hand, a no-training control group is needed, as we have used in the past (Bernstein et al., 2013) to verify that training and not repeated CVCVC consonant identification testing was responsible for the reliable improvement in consonant identification scores.

Scores on the paired-associates training and test tasks generally dropped across lists. This was likely due to interference from previous words or fribbles, and/or fatigue, inasmuch as the three lists were generally undertaken within a single session. Also, there were six training trials per word for the first list and only four trials per word for the second and third lists, likely further contributing to poor performance on later lists.

The Experiment-1 results suggest that visual stimuli during AV training can be an impediment to immediate auditory speech perceptual learning for pre-/perilingually deafened adults with late-acquired cochlear implants and reliance on visual speech. In contrast, previously Bernstein et al. (2013) reported that normal-hearing participants who received only AV paired-associates training, with sinewave vocoded audio, were significantly more successful when AO testing followed than those who received only AO training. It was suggested that the normal-hearing participants used the concordance between visual speech and vocoded audio to learn the novel features of the audio. Also, previous results with normal-hearing participants showed that the medial consonants were most accurately identified in CVCVC stimuli,

whereas in Experiment 1, the initial consonants were most accurately identified.

One explanation for why the results in Experiment 1 were so different from those in Bernstein et al. (2013) is that the cochlear implant participants brought to the perceptual learning task different perceptual abilities and biases, in particular, reliance on visual information and enhanced lipreading ability. A second possibility is that the Experiment-1 protocol here differed in some important way from Experiment 1 in the previous report. This alternative gains some credence given that in Bernstein et al. (2013) there were also discrepancies between their Experiment-1 cross-subjects design and their Experiment 3 within-subjects design for which training alternated list-by-list between AV and AO stimuli. With that design, AV training was associated with overall reduced AO test scores, however there were somewhat fewer training trials with the AV training, inasmuch as training was to criterion rather than with fixed numbers of trials (see Bernstein et al. for a discussion of how the alternation might have led to greater reliance on visual speech for learning).

In Experiment 2 here, normal-hearing participants carried out the same protocol as in Experiment 1 in order to determine whether the Experiment-1 pattern of results could be attributable to participant characteristics and/or to the paradigm itself. A drop in AO paired associates test scores following AV training with normal-hearing participants would support the interpretation that aspects of the paradigm itself resulted in biasing attention to the visual information and thereby impeding auditory speech perceptual learning. In addition, if the normal-hearing participants, like the cochlear implant users, focused on the initial consonant during CVCVC phoneme identification, the implication would be strengthened that the paradigm itself biases what is learned. In fact, quite different results were obtained across Experiments 1 and 2, supporting the general conclusion that the two groups of trainees brought far different perceptual abilities or biases to the training paradigm.

EXPERIMENT 2: NORMAL-HEARING PARTICIPANTS

The acoustic stimuli for Experiment 2 were generated using a custom realtime hardware/software sinusoidal vocoder (Iverson et al., 1998). Frequently, simulation of cochlear implants is carried out using noise-band vocoding (Shannon et al., 1995), which uses speech-derived amplitude modulation of noise bands, but noise-band and sinusoidal vocoding have been compared and shown to deliver similar results (Dorman et al., 1997). The vocoded speech here used 15 filters to amplitude modulate single sinusoids at the center frequencies of each filter, resulting in greatly degraded speech (see below for a more complete description). When consonant identification was tested previously using the same CVCVC stimuli used here, pre-training test scores were approximately 30% correct for initial consonants (Bernstein et al., 2013), a similar level of accuracy to that for cochlear implant patients in Experiment 1.

Normal-hearing participants

Individuals were screened for American English as a first language, normal or corrected-to-normal vision in each eye of 20/30 or better (using a Snellen chart). Normal-hearing participants were screened for normal hearing (25 dB HL or better in each ear for frequencies between 125 Hz-8 KHz, using an Audiometrics GSI 16 audiometer with insert earphones). All 43 of the participants received a lipreading screening test (Auer and Bernstein, 2007). Normal-hearing participants ranged in age from 18 to 49 years (mean = 24.9), with 16 males. The experiment was carried out at the House Research Institute. All participants were paid $12 per hour plus any travel expenses incurred. Participants gave written consent. Human subject participation was approved by the St. Vincent's Hospital Institutional Review Board (Los Angeles, CA).

Stimuli

The stimulus materials were the same as in Experiment 1 but the acoustic stimuli were processed by passing them through a custom realtime hardware/software vocoder (Iverson et al., 1998). The vocoder detected speech energy in thirteen 120-Hz-bandwidth bandpass filters with center frequencies every 150 Hz from 825 Hz through 2625 Hz. Two additional filters were used to convey high frequencies. One was a bandpass filter centered at 3115 Hz with 350 Hz bandwidth and the other a highpass filter with 3565 Hz cutoff. The energy detected in each band was used to amplitude-modulate a fixed-frequency sinewave at the center frequency of that band (and at 3565 Hz in the case of the highpass filter). The sum of the 15 sinewaves comprised the vocoded acoustic signal. This acoustic transformation retained the gross spectral-temporal amplitude information in the waveform while eliminating finer distinctions such as fundamental frequency variations and the natural spectral tilt of the vocal tract resonances. **Figure 7** compares /bɛ/ and /fɛ/ between the original recordings and the vocoded versions.

Apparatus

The testing apparatus in Experiment 2 was the same as in Experiment 1, except that the acoustic waveforms were vocoded in real time rather than processed through a cochlear implant.

Procedure

Other than the acoustic stimuli, the normal-hearing participants received the same protocol as the participants with cochlear implants.

RESULTS

Lipreading scores

The lipreading screening scores were compared across the two training modality assignments (AO-AV, AV-AO) to assure that the two groups were not different, $t_{(40)} = 1.478$, $p = 0.147$. Lipreading scores were also compared across Experiments 1 and 2, and were different, $t_{(68)} = 7.582$, $p = 0.000$. The mean normal-hearing participant's score was 8.1% correct, and the mean cochlear implant user's score was 39.4% correct.

Paired-associates training results

Figure 3B shows the time series of training and test scores in Experiment 2. Examination of **Figure 3A** vs. **3B** suggests that both participant groups began learning a list at roughly the same level of accuracy, but normal-hearing participants were much more accurate by the time training was completed on each list. Also,

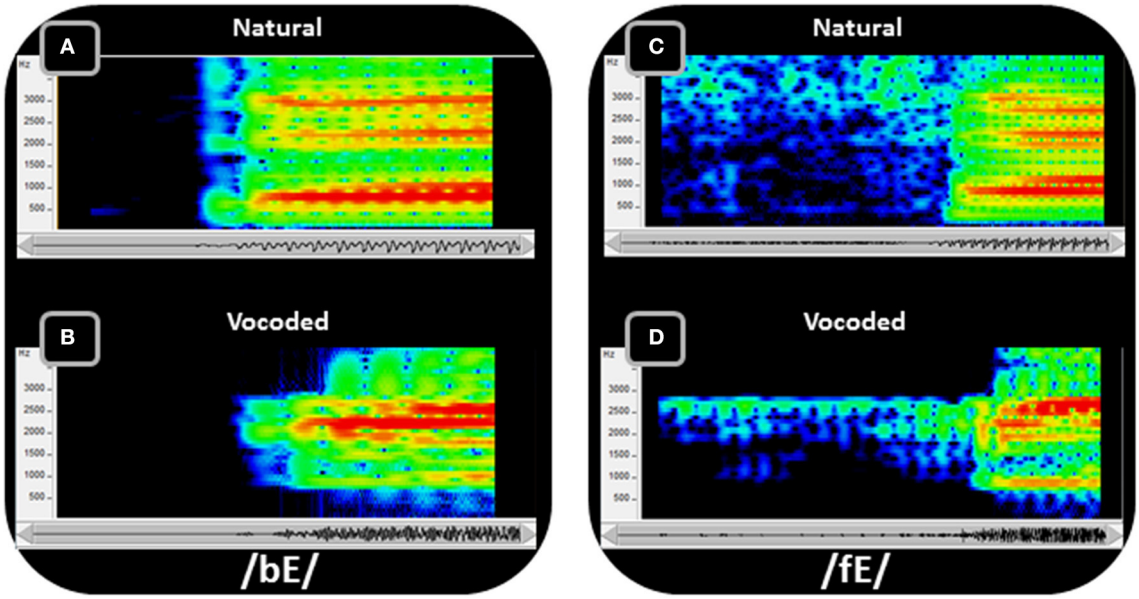


FIGURE 7 | Spectrograms of speech show the concentrations of energy in the spectra over time. Two speech tokens, /bE/ **(A,B)** and /fE/ **(C,D)** (i.e., the vowel in "bet"), are shown in spectrograms of the natural **(A,C)** recorded speech and the vocoded **(B,D)** speech. The frequency range of the spectrograms is limited to 4 kHz, because all of the energy from the vocoder is similarly limited. The amplitudes are represented as a heat map, with red the highest amplitude and dark blue the lowest. In addition to representing the speech as the sum of sinewaves at the center of each vocoder filter (see text), the vocoder also tilted the spectrum so that it did not roll off at approximately 6 dB/octave, which is natural to speech. Thus, the amplitudes of the frequencies vary between the natural and the vocoded speech, in addition to the frequency ranges and spectral detail (adapted from Bernstein et al., 2013).

the pattern of reduced AO test scores following AV training is not present in **Figure 3B**.

Analyses of the training results were carried out using the last training block per list. The within-subjects factors were training list (3) and training period (Period 1: first three lists; Period 2: second 3 lists), and the between-subjects factor was modality assignment (AO-AV, AV-AO). List was a reliable main effect, $F_{(2,\ 40)} = 6.043$, $p = 0.005$, $\eta_p^2 = 0.232$. Scores dropped from List 1 (89.9% correct untransformed) to List 2 (84.0% correct), $F_{(1,\ 41)} = 12.245$, $p = 0.001$, and rebounded somewhat for List 3 (86.5% correct). Training period was also a reliable factor, $F_{(1,\ 41)} = 14.907$, $p = 0.000$, $\eta_p^2 = 0.267$ (Period 1 mean 83.5% correct; Period 2 mean 90.1% correct). The training period by modality assignment interaction was marginally reliable, $F_{(1,\ 41)} = 3.427$, $p = 0.071$, $\eta_p^2 = 0.077$. Period-1 AV training scores (84.5% correct) were somewhat higher than Period-1 AO training scores (83.6% correct), and a slight advantage for AV training continued for Period 2.

Paired-associates AO test results

Following each training list, participants were tested AO on their paired-associates learning for 6 out of the 12 trained associations. The test scores were submitted to analyses with the within subjects factors test list (3) and testing period (Period 1, first 3 lists; Period 2, second 3 lists), and the between subjects factor modality assignment (AO-AV, AV-AO). The main effect of testing period was the only one that was reliable, $F_{(1,\ 41)} = 6.576$, $p = 0.014$, $\eta_p^2 = 0.138$. Means were 85.0% correct over Period 1 tests and 90.2% correct over Period 2 tests.

Paired-associates training and test scores compared

There was no evidence across the full set of participants in Experiment 2 for a change from training to testing as a function of the training modality. However, because this null finding contradicted our earlier results (Bernstein et al., 2013), we carried out a more detailed set of analyses (see further below).

CVCVC forced-choice consonant identification

Participants identified the three consonants in CVCVC stimuli before their first training period (pre-1), after their first training period (post-1), and after their second training period (post-2). The proportion correct scores were computed separately for each consonant position (initial, medial, final) and each test period (pre-1, post-1, and post-2). These scores were submitted to analyses for within subjects factors CVCVC testing period (3) and consonant position (3), and for the between subjects factor training modality (AV-AO, AO-AV). **Table 2** shows the mean scores across positions, testing periods, and modality-assignment groups.

Reliable effects were obtained for test period, $F_{(2,\ 39)} = 129.811$, $p = 0.000$, $\eta_p^2 = 0.869$, and consonant position, $F_{(2,\ 39)} = 171.216$, $p = 0.000$, $\eta_p^2 = 0.898$, and their interaction $F_{(4,\ 37)} = 2.629$, $p = 0.050$, $\eta_p^2 = 0.221$. However, none of the simple tests explained the interaction. Scores from post-1 were higher than pre-1, $F_{(1,\ 40)} = 75.286$, $p = 0.000$, $\eta_p^2 = 0.653$, and post-2 were higher than post-1, $F_{(1,\ 40)} = 13.664$, $p = 0.001$, $\eta_p^2 = 0.255$ (pre-1, 35.8%; post-1, 46.9%; post-2, 51.1%). The scores for the medial consonant position were higher than for either the initial $F_{(1,\ 40)} = 260.202$, $p = 0.000$, $\eta_p^2 = 0.867$, or the

final position, $F_{(1,\ 40)} = 264.936$, $p = 0.000$, $\eta_p^2 = 0.869$ (initial, 36.9%; medial, 56.2%; final, 40.7%).

Paired-associates results in relationship to previous findings

We reported previously using a similar training paradigm with normal-hearing participants that AV training can be more effective than AO training for AO learning (Experiment 1, Bernstein et al., 2013). In that study, in a between subjects design, participants were assigned to AV or AO training. Training was on four lists of 12 paired-associates, with three training blocks per list. Analyses were performed using the data from participants who scored 75% or greater by the third training block on all four lists, and mean training scores were 94% correct. Here in Experiment 2, Period-1 training was essentially a between-subjects design, so results were further probed for evidence from Period 1 that auditory perceptual learning was greater with AV training. There were, as already noted however, several potentially important differences between the current Experiment 2 and the previously published study. Here, only List 1 was trained for three blocks. Only two training blocks were given for Lists 2 and 3, and training scores dropped reliably across lists.

When the criterion of 75% correct on final blocks was imposed for inclusion of Experiment 2 participants' data, sample sizes for the AO-AV assignment were reduced from 22 to 17 and for the AV-AO assignment from 21 to 15 participants. Mean training scores were 89.9% for AO-trained and 93.1 for AV-trained participants. Thus, as with the participants in Experiment 1, there was an AV advantage during training, $F_{(1,\ 30)} = 4.971$, $p = 0.033$, $\eta_p^2 = 0.142$. But unlike the outcome for the cochlear implant users, there was an AV training advantage for AO test scores, albeit at a marginal level of reliability, $F_{(1,\ 30)} = 3.229$, $p = 0.082$, $\eta_p^2 = 0.097$ (observed power = 0.413). The AV-trained participants scored mean 80.1% on AO tests, and AO-trained participants scored mean 77.6%. These means contrasted with the previous study for which AV-trained mean scores were 97% correct and AO-trained means scores were 92% correct. The higher scores obtained previously are likely attributable to longer training on more lists and training on only one list per day.

Discussion

Results of Experiment 2 showed that normal-hearing participants did learn differently than did the cochlear implant users in Experiment 1. Normal-hearing participants' test scores did not drop following AV training, and there was evidence that AV training was superior to AO training in terms of AO paired-associates test scores. In comparison with the previous study with normal-hearing adults (Bernstein et al., 2013), less training was given on fewer lists, and these task differences across experiments were likely responsible for the less reliable AV training advantage in Experiment 2 and the generally lower scores.

Across Experiments 1 and 2, the pattern of CVCVC phoneme identification scores was clearly different. Cochlear implant participants were most accurate for the first consonant in the CVCVC phoneme identification stimuli, and normal-hearing participants were most accurate for the medial consonant. Interestingly, phoneme scores across groups were similar for the initial consonant, a point revisited below.

GENERAL DISCUSSION

Our environment affords multisensory stimulation that is integrated during perception. The possibility that information obtained through one sensory system can assist perceptual learning by a different sensory system has apparent face validity (Merabet et al., 2005). But neuroplastic changes associated with loss or a disorder of a sensory system could result in functional system modifications that instead are impediments to perceptual learning under multisensory conditions. In postlingually deafened adults, whose sensory systems developed normally followed by auditory loss and then restoration, concordant visual speech could be very useful for learning to perceive auditory input from a cochlear implant (Rouger et al., 2007). But pre-/perilingually deafened individuals who acquire cochlear implants late were never normally stimulated (Gilley et al., 2010; Kral and Sharma, 2012), and an early visual dominance could lead to a long-lasting bias in sensory processing and organization toward the dominant visual modality. In this study, the mean normal-hearing participant's lipreading screening score was 8.1% correct, and the mean cochlear implant user's score was 39.4% correct, supporting the point that they brought different perceptual abilities to the experiments.

This study was carried out to learn how training with audiovisual speech stimuli affects auditory speech perceptual learning in prelingually deafened adults with late-acquired cochlear implants (Experiment 1) in comparison with normal-hearing adults (Experiment 2). Training used a paired-associates paradigm in which participants learned to associate twelve spoken CVCVC non-sense words with 12 fribble non-sense pictures. Six lists were trained, and training on the first three lists commenced with either AV or AO stimuli (Period 1); then training continued with the opposite training modality for three lists (Period 2). AO learning for each list of stimuli was tested immediately after training. A CVCVC phoneme identification task was administered with untrained stimuli before paired-associates training and testing, after Period 1, and after Period 2. Participants identified each of the consonants in the non-sense words.

In Experiment 1, prelingually deafened adults with late-acquired cochlear implants were able to learn the paired-associates, and their AO test scores improved 7.3 percentage points between Periods 1 and 2. Also, consonant identification for the consonants in untrained CVCVC stimuli improved between the second and third administrations of consonant identification testing, with a reliable mean improvement of 3.5 percentage points. Initial consonants were most accurately identified. The results on the phoneme identification task suggest the possibility that participants learned sub-lexical auditory speech features during the paired-associates task, even though no feedback or explicit training of consonant identification was provided. However, as noted above, a no-training control is needed to confirm that the improvement in consonant identification was indeed due to the paired-associates training.

The answer to the main question of how modality of training affects auditory perceptual learning in these cochlear implant users was shown in terms of the AO paired-associates test scores and their relationship to training scores. During *Period 1* of training, cochlear implant users' training scores were similar

independent of training modality (AO or AV), suggesting that perceptual modality *per se* did not control learning the paired-associates task or the associations. However, large group differences emerged in the comparison between training and AO test scores, with AV-trained participants' AO test scores lower than training scores by an average 22.8 percentage points; while the AO-trained participants' scores stayed essentially the same at test (1.8 percentage points different between training and testing). During *Period 2*, again a large drop between AV training and AO test scores (11.5 percentage points) was observed. Overall, these results suggest that visual speech impeded auditory paired-associates learning.

Experiment 2 investigated whether the results in Experiment 1 were attributable to the type of participant in Experiment 1, or to how the paradigm was administered, inasmuch as previous evidence suggested that the paradigm itself can influence whether auditory perceptual learning takes place (Bernstein et al., 2013). The results with normal-hearing participants were dramatically different from those in Experiment 1: AO test scores did not decline and even benefited following AV paired-associates training. When the results were analyzed to determine whether previous ones showing AV benefit (Bernstein et al., 2013) had been replicated, AV training in Experiment 2 was shown to be more effective than AO training, albeit at a reduced level of statistical reliability ($p = 0.082$), which was attributed to the truncated training protocol relative to that of the previous study. It could also be the case that the normal-hearing participants here paid less attention to the visual stimuli. Future use of eye tracking is needed to determine whether learning is related to different gaze patterns.

There was no ambiguity about whether there was a difference in learning patterns between Experiments 1 and 2 here. On a per-list basis, the adults with cochlear implants always had better AV training scores than AO test scores. In contrast, normal-hearing adults maintained their performance levels or were more successful during AO testing when it followed AV training.

In addition, the CVCVC consonant identification scores of normal-hearing participants improved across the three test periods. But they identified the medial consonant most accurately: The cochlear implant users were more accurate for the initial consonant. Notably, normal-hearing and cochlear implant participants had similar scores for the initial consonant of the CVCVC identification stimuli, suggesting the possibility that visual bias on the part of the cochlear implant users limited access to available auditory information. We return to these points below.

MULTISENSORY REVERSE HIERARCHY THEORY

The results reported here support the conclusion that perceptual learning within a habilitated sensory system following life-long sensory deprivation requires more than afferent activation by a sensory prosthetic device. Evidence on the effects of deafness on subcortical auditory system and primary auditory cortex suggests that *ceteris paribus* late-implantation in this patient population could be more successful than it typically is (for reviews see Kral and Eggermont, 2007; Kral and Sharma, 2012): However, the evidence suggests that even with neuroplastic changes following cochlear implant habilitation, corticofugal influences are likely deficient. The role of top-down connections and processing should be taken into account in theorizing about and crafting approaches that facilitate perceptual learning. A severely limiting factor for auditory perceptual learning in pre-/perilingually deafened adults with late-acquired cochlear implants is likely their reduced representations of high-level auditory speech categories such as phoneme categories (Kral and Eggermont, 2007), coupled with their enhanced ability to lipread. The critical need for high-level representations to guide lower-level auditory perceptual learning is explained within so-called *reverse hierarchy theory* (RHT) (Ahissar and Hochstein, 1997; Ahissar et al., 2008).

The *hierarchy* in *RHT* refers to the cortical organization of sensory-perceptual pathways (Felleman and Van Essen, 1991; Kaas and Hackett, 2000; Kral and Eggermont, 2007). Although pathways are not strictly hierarchical, their organization is such that higher cortical levels typically show selectivity for increasingly complex stimuli combined with an increasing tolerance to stimulus transformation and increasing response to perceptual category differences (Hubel and Wiesel, 1962; Ungerleider and Haxby, 1994; Logothetis and Sheinberg, 1996; Binder et al., 2000; Zeki, 2005; Obleser et al., 2007).

According to RHT (Ahissar and Hochstein, 1997; Kral and Eggermont, 2007; Ahissar et al., 2008), immediate perception relies on already-established higher-level representations in the bottom-up sensory-perceptual pathway. When a *new* perceptual task needs to be carried out, naïve performance is initiated on the basis of immediately available high-level perception. However, if the task cannot be readily performed with the existing mapping of lower-level to higher-level representations, and/or if there is incentive to increase the efficiency of task performance, then perceptual learning can occur. According to RHT, perceptual learning is by definition the access to and remapping of lower-level input representations to higher-level representations. Thus, perceptual learning involves dissimilar lower-level input representations being remapped to the same higher-level representations, or similar lower-level input representations being remapped to different higher-level representations.

However, RHT also posits that perceptual learning requires "perception with scrutiny." That is, a backward (top-down) search from a higher level of the representational hierarchy must be initiated to access lower-level representations. A more effective forward mapping can then be made in terms of altered convergence and/or divergence patterns within existing neural networks (Jiang et al., 2007; Kral and Eggermont, 2007; Ahissar et al., 2008).

NEURAL RESOURCES FOR MULTISENSORY RHT

The results of this study and previous studies suggest that adults with normal hearing are able to use visual stimuli to direct/improve scrutiny of auditory speech features in order to learn vocoded speech features (Wayne and Johnsrude, 2012; Bernstein et al., 2013). Multisensory RHT extends RHT to perceptual learning initiated through scrutiny of features in one sensory system's representations being initiated by another system's representations (Bernstein et al., 2013). In order for such scrutiny to be possible, there must be neural connections available across sensory systems. Many results point to multisensory integration at higher cortical levels, particularly the posterior

superior temporal sulcus with potential for feedback to lower level cortices (e.g., Miller and d'Esposito, 2005; Hasson et al., 2007; Bernstein et al., 2008; Nath and Beauchamp, 2011). The evidence is extensive on the sheer diversity and extent of cortical and subcortical multisensory connections (e.g., Foxe and Schroeder, 2005; Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008; Kayser et al., 2012). Thus, the neural resources are available for higher-level representations in one sensory-perceptual system to gain access to lower-level representations in a different sensory-perceptual system, as well as for low-level cross-sensory connections to activate early areas (Ghazanfar et al., 2008; Falchier et al., 2012).

PERCEPTION WITHOUT SCRUTINY

According to multisensory RHT, when auditory speech features are novel to the naïve listener, as is noise or sinewave-vocoded speech to normal-hearing listeners, or when auditory speech features have not been adequately learned by cochlear implant users, familiar concurrent visual speech features can compensate through immediate high-level perception. Importantly, compensation could arise in more than one way. The visual information might be sufficient by itself to carry out the task, completely obviating the need for auditory input. Or the familiar concurrent visual information might combine with deficient auditory information (Sumby and Pollack, 1954; Summerfield, 1987; Ross et al., 2007). In either case there may be no need for perception with scrutiny, and auditory perceptual learning is not expected. In this study, the paired-associates learning task can be carried out accurately on the basis of the visual stimuli only, and in an ongoing study (Eberhardt et al., submitted), we have shown that normal-hearing adults can do so. Thus, the cochlear-implant users here with their enhanced visual speech perception could have relied entirely on visual speech and/or combined visual and auditory features to carry out the paired-associates learning task without additional scrutiny of the auditory stimuli. This type of perception without scrutiny would predictably result in a steep drop in AO test scores when the visual stimuli were not shown, as occurred here. Had the visual stimuli here been less visually distinct, it is possible that cochlear implant users might have relied less on the visual stimuli during AV training, which suggests that AV stimuli could be developed to be better promoters of auditory perceptual learning. It is also possible that the between-subjects design here, which incorporated cross-over between AV and AO training modalities, was itself important to learning. Control experiments with only AV or only AO training are needed to ascertain definitively whether or not cochlear implant participants benefit from being trained in only one vs. both conditions.

CONCURRENT VISUAL AND AUDITORY SPEECH FEATURES

The validity of the suggestion that visual speech can guide auditory perceptual learning depends on visual speech being adequately informative. Visual speech stimuli are however frequently characterized as limited in speech information. For example, the so-called *viseme*, that is, groupings of visually confusable phonemes, such as "b," "p," and "m," are sometimes said to be perceptually indistinguishable (Massaro et al., 2012). But discrimination can be reliable for phonemes within visemes (Files et al., 2013; Files et al., in preparation).

In addition, visual speech information is highly distributed across the cheeks, lips, jaw, and tongue (when it can be glimpsed inside the mouth opening), and the motions of these structures are in highly predictable relationships with auditory speech information (Jackson et al., 1976; Yehia et al., 1998; Jiang et al., 2002; Jiang and Bernstein, 2011). Furthermore, normal-hearing adults systematically perceive the concurrence/congruity of auditory and visual speech when the stimuli are mismatched (Jiang and Bernstein, 2011), consistent with findings of visual speech representations in the high level vision pathway (Bernstein et al., 2011; Files et al., 2013).

The cochlear implant users here appear to approach the paired-associates learning task with possibly limited ability to use concordance across auditory and visual representations in order to learn the auditory speech information and also appear to even carry over patterns of perceptual attention for visual speech into auditory perception. Cochlear implant users were most accurate for initial consonants in CVCVCs, and normal-hearing adults most accurate for medial consonants (see **Table 2**). Participants with cochlear implants had initial consonant correct scores (33.7%) that were higher than medial scores (28.9%). Normal-hearing participants' consonant position scores were most accurate for medial consonants (i.e., initial 36.9%; medial, 56.2%; and final, 40.7%). Lipreaders are most accurate for initial consonants, and this is true whether they are deaf or hearing (Auer and Bernstein, in preparation). Apparently, the initial consonant affords the most information to the lipreader. However, auditory perception can be more accurate for medial consonants, because in the VCV position, consonant information is distributed across the preceding and following vowel transitions (Stevens, 1998). Intriguingly, here, initial consonant scores were similar across normal-hearing and cochlear implant participants, suggesting the possibility that having a visual speech bias is an impediment to learning available auditory speech features even under auditory-only training conditions. Training that focused on medial consonants might be very effective for cochlear implant users, but training programs typically use monosyllabic syllables or words (e.g., Fu and Galvin, 2007, 2008).

Feedback based on orthography can also be used by normal-hearing individuals to learn novel acoustic speech stimuli (Hervais-Adelman et al., 2008). But visual speech presented in its normal temporal relationship with auditory speech has the advantages of being closely aligned in time, displaying similar internal temporal dynamics (i.e., the vocal tract actions that produce acoustic speech signals are the same actions that produce optical ones), and of already being tightly processed with auditory speech. The problem then for the pre-/perilingually deafened individual with a late-acquired cochlear implant is to use the available audiovisual stimulus concordance to discern new auditory information and not evade auditory perceptual learning.

IMPLICATIONS FOR FUTURE RESEARCH

Only a small minority (about 10%) of individuals with pre-/perilingual deafness have deaf parents who communicate

with sign language (https://www.nidcd.nih.gov/StaticResources/health/healthyhearing/tools/pdf/commoptionschild.pdf). Thus, the vast majority of deaf children encounter spoken language daily, and as a group they become as adults better lipreaders than normal-hearing individuals (Bernstein et al., 2000, 2001; Mohammed et al., 2006; Auer and Bernstein, 2007; Kyle et al., 2013). If pre-/perilingually deafened individuals do not acquire a cochlear implant, they frequently do use high power hearing aids. The stimulation from the hearing aids likely is mostly low frequencies that can represent the voice fundamental frequency and can also be perceived via somatosensory stimulation through mechanical vibration (Nober, 1967; Boothroyd and Cawkwell, 1970; Bernstein et al., 1998), which can be associated with increased vibrotactile activation of auditory cortices (Auer et al., 2007; Karns et al., 2012). Thus, deafness is associated with neuroplastic changes involving both somatosensory and visual stimulation. However, low-frequency speech information associated with the voice fundamental frequency can only provide a highly reduced representation of speech that would be most effective in combination with visual stimuli and again would bias individuals with late-acquired implants away from use of segmental auditory information.

Vocoded speech has been used with normal-hearing participants to simulate auditory perception with a cochlear implant and to model learning (Faulkner et al., 2000; Fu and Galvin, 2007; Wayne and Johnsrude, 2012). Along with the results on neuroplasticity, the present study demonstrates that the quality of the speech input may be a necessary but is certainly not a sufficient condition for simulating effects with a cochlear implant in pre-/perilingually deafened late-implanted adults. Valid simulation would seem to require also accounting for the perceptual enhancements or biases that the pre-/perilingually deafened individual brings to the perceptual learning task. In the case of simulating pre-/perilingually deafened adults, strong pre-existing perceptual biases in one sensory system that need to be overcome through training of another need to be simulated. However, because these "biases" observed in pre-/perilingually deafened adults are likely supported by neuroplastic changes, such as recruitment of auditory cortical areas by vision (Finney et al., 2001; Karns et al., 2012; Bottari et al., 2014), they are *per se* unlikely to be simulable in normal-hearing adults. Simulation of post-lingually deafened implant users might seem more valid, because their initial perceptual development established a normal relationship between auditory and visual perception. However, even in these adults, there is evidence for a reliance on vision not present in normal-hearing adults (Rouger et al., 2007, 2008).

We have recently approached the issue of trainees' primary modality for speech perception by carrying out experiments with normal-hearing adults who were trained with the paired-associates paradigm used here and the training goal to learn visual speech stimuli (i.e., to learn to lipread) (Eberhardt et al., submitted). In that study, vocoded acoustic speech impeded visual-only learning, but vibrotactile vocoded speech promoted learning. Thus, our recent results underscore the potential importance of the trainee's primary speech perceptual modality during training.

IMPLICATIONS FOR TRAINING

Overall, the results here could be interpreted as strong support for training under only auditory conditions or for reducing the clarity of visual speech stimuli to focus attention on the auditory stimuli (Huyse et al., 2013). But either of those options would reduce the ability to use the concordance between auditory and visual speech features to access potentially useful auditory features.

An alternative approach might be to use artificial visual or vibrotactile stimuli to target auditory feature distinctions. For example, we have shown that non-speech stimuli, such as a picture of a square and/or a vibrotactile buzz can enhance the efficiency to detect an auditory speech signal in noise (Bernstein et al., 2004; Tjan et al., 2014). Novel non-speech visual or vibrotactile stimuli that correlate with to-be-learned speech features might be useful for training, because they would not be available to the naïve perceiver as a substitute for speech information. Another type of concordant stimuli that is already used in training deaf children is cued speech (Cornett, 1967; Aparicio et al., 2012). Cued speech uses a small number of manual cues to disambiguate difficult visual speech stimuli and has been shown to be highly effective in establishing normal phonological representations. A cuing system based on disambiguating auditory features and designed for cochlear implant users might be useful in training.

In a study of the McGurk effect (McGurk and MacDonald, 1976) with cochlear implant and normal-hearing children, Huyse et al. (2013) showed that by reducing the clarity of visual speech information in blocks that included AO, AV and visual-only stimuli, AO scores improved. The authors speculated that the unreliable visual information in the mixed context of VO, AO, and AV stimuli led to a shift in attention to the auditory input. Above, it was suggested that training with a less visually distinct word set could promote better use of audiovisual concordance. Reduction in visual clarity would however reduce concordant information and could lead alternatively to less effective training.

Johnsrude and colleagues have carried out a number of AO training experiments on vocoded speech with normal-hearing adults (Davis et al., 2005; Hervais-Adelman et al., 2008, 2011). Their experiments show that the organization of individual training trials influences learning, with learning the vocoded speech enhanced by knowing the words in sentences and then hearing the degraded speech. Unfortunately, with pre-/perilingual deafness, orthographic feedback for the lexical content of stimuli may not be as effective, because reading levels are reduced in this group (Trybus and Karchmer, 1977; King and Quigley, 1985; Allen, 1986), and obviously clear speech is not available for feedback. Another alternative would be to present AO, then AV, then AO stimuli to possibly encourage using visual and/or audiovisual representations to access auditory features when the visual are removed within the same training trial (Wayne and Johnsrude, 2012).

A cochlear implant for pre-/perilingually deafened adults could be useful in acquiring new vocabulary, as these individuals tend to lag behind normal-hearing adults in terms of reading ability and vocabulary (Aparicio et al., 2012). The results here suggest that AV training could be very effective for learning new

words and their semantic relationships, which is itself a valid goal for enhancing speech understanding. Lexical processes have been implied in the promotion of perceptual learning in normal-hearing adults across levels of speech (Davis et al., 2005; Davis and Johnsrude, 2007; Ahissar et al., 2008; Samuel and Kraljic, 2009; Bernstein et al., 2013). Furthermore, as suggested earlier, we expect that there is a positive feedback relationship between learning new vocabulary and perceptual learning of auditory input: With greater knowledge of the lexicon comes more opportunities to use top-down processes to guide discernment of auditory input (Kral and Eggermont, 2007). In addition, lexical knowledge appears to be a pre-requisite for certain types of automatic perceptual adjustments to ambiguous auditory speech stimuli, referred to as *perceptual learning* or *recalibration* elsewhere in the literature (Samuel and Lieblich, 2014). While cochlear implant training frequently uses words in training tasks designed to contrast specific phonemes or features, and positive results are attributed implicitly or explicitly to the focused contrast learning at the sub-lexical level (e.g., Fu and Galvin, 2007, 2008), it could as well be the case that lexical effects operate separately from the effects of structured stimulus contrasts (Samuel and Lieblich, 2014).

CONCLUSIONS

In this study, training improved auditory-only test performance for paired-associates and for untrained consonant identification in CVCVC non-sense words. However, training with AV vs. AO speech resulted in a different pattern of performance in cochlear implant users with late-acquired implants vs. normal-hearing adults. In the cochlear implant users, AV training was followed by steep declines in AO test scores; while AV training was followed by stable or even somewhat higher test scores in normal-hearing adults. The contrast across listener groups suggests that they bring to the task perceptual differences that can bias learning. Pre-/perilingually deaf adults have experienced a lifetime of reliance on vision, which may lead them to rely on the visual part of audiovisual stimuli during training. Indeed the cochlear implant users here had much higher lipreading ability than the normal-hearing participants. Multisensory reverse hierarchy theory suggests that in order to use visual speech for auditory perceptual learning, the concordance between auditory and visual speech stimuli must be used to discern and remap available auditory input rather than combine whatever auditory speech has already been learned with readily available visual information. While the inference might be taken from this study that auditory-only training for cochlear implant users should remove the potential to substitute knowledge of visual speech for learning auditory features, reverse hierarchy theory also suggests that auditory-only training would preclude access to important concordant visual information that could guide attention to available lower-level auditory input speech features. That the lower-level information is indeed available is implied by the similarity across normal-hearing and cochlear implant participants in their accuracies for initial consonants with CVCVC stimuli vs. the discrepancy across groups for medial consonants (much higher scores on the part of the normal-hearing). Given similar initial consonant accuracies across groups, the implant users' poorer medial consonant performance appears to be limited at least in part by their perceptual biases, not by their auditory input processing. Indeed, attention to initial consonants is reminiscent of the pattern observed in deaf and hearing lipreaders. Biased attention to initial consonants could limit acquiring additional *available* auditory information from consonants in intervocalic positions. A comprehensive view of language use also suggests that audiovisual training has an important role for vocabulary learning, and that vocabulary growth can in turn promote perceptual learning. This study also highlights a serious pitfall for research that attempts to simulate cochlear implant use with normal-hearing adults, specifically, that results need not generalize to actual cochlear implant users who have far different perceptual experience than normal-hearing adults. Additional studies are needed to understand how individual perceptual experience across the lifespan influences perceptual learning.

ACKNOWLEDGMENTS

This study was carried out with the support of NIH (DC008308, Bernstein PI). We thank research assistants, technicians, and participants, without whom this study could not have been carried out.

REFERENCES

Ahissar, M., and Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature* 387, 401–406. doi: 10.1038/387401a0

Ahissar, M., Nahum, M., Nelken, I., and Hochstein, S. (2008). Reverse hierarchies and sensory learning. *Philos. Trans. R. Soc. B* 364, 285–299. doi: 10.1098/rstb.2008.0253

Allen, T. E. (1986). "Patterns of academic achievement amoung hearing impaired students: 1974–1983," in *Deaf Children in America*, eds A. N. Schildroth and M. A. Karchmer (San Diego, CA: College-Hill Press), 161–206.

Aparicio, M., Peigneux, P., Charlier, B., Neyrat, C., and Leybaert, J. (2012). Early experience of Cued Speech enhances speechreading performance in deaf. *Scand. J. Psychol.* 53, 41–46. doi: 10.1111/j.1467-9450.2011.00919.x

Auer, E. T. Jr., and Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *J. Acoust. Soc. Am.* 102, 3704–3710. doi: 10.1121/1.420402

Auer, E. T. Jr., and Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *J. Speech Lang. Hear. Res.* 50, 1157–1165. doi: 10.1044/1092-4388(2007/080)

Auer, E. T. Jr., Bernstein, L. E., Sungkarat, W., and Singh, M. (2007). Vibrotactile activation of the auditory cortices in deaf versus hearing adults. *Neuroreport* 18, 645–648. doi: 10.1097/WNR.0b013e3280d943b9

Bernstein, L. E., Auer, E. T. Jr., Eberhardt, S. P., and Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Front. Neurosci.* 7:34. doi: 10.3389/fnins.2013.00034

Bernstein, L. E., Auer, E. T. Jr., and Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Commun.* 44, 5–18. doi: 10.1016/j.specom.2004.10.011

Bernstein, L. E., Auer, E. T. Jr., and Tucker, P. E. (2001). Enhanced speechreading in deaf adults: can short-term training/practice close the gap for hearing adults? *J. Speech Lang. Hear. Res.* 44, 5–18. doi: 10.1044/1092-4388(2001/001)

Bernstein, L. E., Auer, E. T. Jr., Wagner, M., and Ponton, C. W. (2008). Spatio-temporal dynamics of audiovisual speech processing. *Neuroimage* 39, 423–435. doi: 10.1016/j.neuroimage.2007.08.035

Bernstein, L. E., Demorest, M. E., and Tucker, P. E. (2000). Speech perception without hearing. *Percept. Psychophys.* 62, 233–252. doi: 10.3758/BF03205546

Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z. L., and Joshi, A. (2011). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Hum. Brain Mapp.* 32, 1660–1676. doi: 10.1002/hbm.21139

Bernstein, L. E., Tucker, P. E., and Auer, E. T. Jr. (1998). Potential perceptual bases for successful use of a vibrotactile speech perception aid. *Scand. J. Psychol.* 39, 181–186. doi: 10.1111/1467-9450.393076

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cereb. Cortex* 10, 512–528. doi: 10.1093/cercor/10.5.512

Bodmer, D., Shipp, D. B., Ostroff, J. M., Ng, A. H., Stewart, S., Chen, J. M., et al. (2007). A comparison of postcochlear implantation speech scores in an adult population. *Laryngoscope* 117, 1408–1411. doi: 10.1097/MLG.0b013e318068b57e

Boothroyd, A., and Cawkwell, S. (1970). Vibrotactile thresholds in pure tone audiometry. *Acta Otolaryngol.* 69, 381–387. doi: 10.3109/00016487009123382

Bottari, D., Heimler, B., Caclin, A., Dalmolin, A., Giard, M. H., and Pavani, F. (2014). Visual change detection recruits auditory cortices in early deafness. *Neuroimage* 94, 172–184. doi: 10.1016/j.neuroimage.2014.02.031

Cornett, R. O. (1967). Cued Speech. *Am. Ann. Deaf* 112, 3–13.

Davis, M. H., and Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hear. Res.* 229, 132–147. doi: 10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J. Exp. Psychol. Gen.* 134, 222–241. doi: 10.1037/0096-3445.134.2.222

Dettman, S., Wall, E., Constantinescu, G., and Dowell, R. (2013). Communication outcomes for groups of children using cochlear implants enrolled in auditory-verbal, aural-oral, and bilingual-bicultural early intervention programs. *Otol. Neurotol.* 34, 451–459. doi: 10.1097/MAO.0b013e3182839650

Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J. Acoust. Soc. Am.* 102, 2403–2411. doi: 10.1121/1.419603

Driver, J., and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron* 57, 11–23. doi: 10.1016/j.neuron.2007.12.013

Dunn, L. M., and Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised*. Pines, MN: American Guidance Service.

Erber, N. P. (1975). Auditory-visual perception of speech. *J. Speech Hear. Disord.* 40, 481–492. doi: 10.1044/jshd.4004.481

Falchier, A., Cappe, C., Barone, P., and Schroeder, C. E. (2012). "Sensory convergence in low-level cortices," in *The New Handbook of Multisensory Processing*, ed B. E. Stein (Cambridge, MA: MIT), 67–79.

Faulkner, A., Rosen, S., and Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: implications for cochlear implants. *J. Acoust. Soc. Am.* 108, 1877–1887. doi: 10.1121/1.1310667

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Files, B. T., Auer, E. T. Jr., and Bernstein, L. E. (2013). The visual mismatch negativity elicited with visual speech stimuli. *Front. Hum. Neurosci.* 7:371. doi: 10.3389/fnhum.2013.00371

Finney, E. M., Fine, I., and Dobkins, K. R. (2001). Visual stimuli activate auditory cortex in the deaf. *Nat. Neurosci.* 4, 1171–1173. doi: 10.1038/nn763

Foxe, J. J., and Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport* 16, 419–423. doi: 10.1097/00001756-200504040-00001

Fryauf-Bertschy, H., Tyler, R. S., Kelsay, D. M. R., Gantz, B. J., and Woodworth, G. G. (1997). Cochlear implant use by prelingually deafened children: the influences of age at implant and length of device use. *J. Speech Hear. Res.* 40, 183–199. doi: 10.1044/jslhr.4001.183

Fu, Q. J., and Galvin, J. J. 3rd. (2007). Perceptual learning and auditory training in cochlear implant recipients. *Trends Amplif.* 11, 193–205. doi: 10.1177/1084713807301379

Fu, Q. J., and Galvin, J. J. 3rd. (2008). Maximizing cochlear implant patients' performance with advanced speech training procedures. *Hear. Res.* 242, 198–208. doi: 10.1016/j.heares.2007.11.010

Ghazanfar, A. A., Chandrasekaran, C., and Logothetis, N. K. (2008). Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integratn in *Rhesus mondeys*. *J. Neurosci.* 28, 4457–4469. doi: 10.1523/JNEUROSCI.0541-08.2008

Ghazanfar, A. A., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10, 278–285. doi: 10.1016/j.tics.2006.04.008

Gilley, P. M., Sharma, A., Mitchell, T. V., and Dorman, M. F. (2010). The influence of a sensitive period for auditory-visual integration in children with cochlear implants. *Restor. Neurol. Neurosci.* 28, 207–218. doi: 10.3233/RNN-2010-0525

Giraud, A. L., Price, C. J., Graham, J. M., Truy, E., and Frackowiak, S. J. (2001). Cross-modal plasticity underpins language recovery after cochlear implantation. *Neuron* 30, 657–663. doi: 10.1016/S0896-6273(01)00318-X

Hammill, D. D., Pearson, N. A., and Wiederholt, J. L. (1996). *Comprehensive Test of Nonverbal Intelligence*. Austin, TX: Pro-Ed.

Hasson, U., Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56, 1116–1126. doi: 10.1016/j.neuron.2007.09.037

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., and Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. *J. Exp. Psychol. Hum. Perform. Percept.* 34, 460–474. doi: 10.1037/0096-1523.34.2.460

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., Taylor, K. J., and Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *J. Exp. Psychol. Hum. Perform. Percept.* 37, 293–295. doi: 10.1037/a0020772

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Huyse, A., Berthommier, F., and Leybaert, J. (2013). Degradation of labial information modifies audiovisual speech perception in cochlear-implanted children. *Ear Hear.* 34, 110–121. doi: 10.1097/AUD.0b013e3182670993

Iverson, P., Bernstein, L. E., and Auer, E. T. Jr. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Commun.* 26, 45–63. doi: 10.1016/S0167-6393(98)00049-1

Jackson, P. L., Montgomery, A. A., and Binnie, C. A. (1976). Perceptual dimensions underlying vowel lipreading performance. *J. Speech Hear. Res.* 19, 796–812. doi: 10.1044/jshr.1904.796

Jiang, J., Alwan, A., Keating, P., Auer, E. T. Jr., and Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP J. Appl. Signal. Process.* 2002, 1174–1188. doi: 10.1155/S1110865702206046

Jiang, J., and Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol. Hum. Perform. Percept.* 37, 1193–1209. doi: 10.1037/a0023100

Jiang, X., Bradley, E. D., Rini, R. A., Zeffiro, T., VanMeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron* 53, 891–903. doi: 10.1016/j.neuron.2007.02.015

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799. doi: 10.1073/pnas.97.22.11793

Karns, C. M., Dow, M. W., and Neville, H. J. (2012). Altered cross-modal processing in the primary auditory cortex of congenitally deaf adults: a visual-somatosensory fMRI study with a double-flass illusion. *J. Neurosci.* 32, 9626–9638. doi: 10.1523/JNEUROSCI.6488-11.2012

Kayser, C., Petkov, C. I., Remedios, R., and Logothetis, N. K. (2012). "Multisensory influences on auditory processing: perspectives from fMRI and electrophysiology," in *The Neural Bases of Multisensory Processes*, eds M. M. Murray and M. T. Wallace (Boca Raton, FL: CRC). Available online at: http://www.ncbi.nlm.nih.gov/books/NBK92843/

King, C. M., and Quigley, S. P. (1985). *Reading and Deafness*. San Diego, CA: College-Hill Press.

Knudsen, E. I. (2004). Sensitive periods in the development of the brain and behavior. *J. Cogn. Neurosci.* 16, 1412–1425. doi: 10.1162/0898929042304796

Kral, A., and Eggermont, J. J. (2007). What's to lose and what's to learn: development under auditory deprivation, cochlear implants and limits of cortical plasticity. *Brain Res. Rev.* 56, 259–269. doi: 10.1016/j.brainresrev.2007.07.021

Kral, A., and Sharma, A. (2012). Developmental neuroplasticity after cochlear implantation. *Trends Neurosci.* 35, 111–122. doi: 10.1016/j.tins.2011.09.004

Kyle, F. E., Campbell, R., Mohammed, T., Coleman, M., and Macsweeney, M. (2013). Speechreading development in deaf and hearing children: introducing the test of child speechreading. *J. Speech Hear. Res.* 56, 416–426. doi: 10.1044/1092-4388(2012/12-0039)

Lamoré, P. J., Huiskamp, T. M., van Son, N. J., Bosman, A. J., and Smoorenburg, G. F. (1998). Auditory, visual and audiovisual perception of segmental speech features by severely hearing-impaired children. *Audiology* 37, 396–419. doi: 10.3109/00206099809072992

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Manrique, M., Cervera-Paz, F. J., Huarte, A., Perez, N., Molina, M., and Garcia-Tapia, R. (1999). Cerebral auditory plasticity and cochlear implants. *Int. J. Pediatr. Otorhinolaryngol.* 49, S193–S197. doi: 10.1016/S0165-5876(99)00159-7

Massaro, D. W., Cohen, M. M., Tabain, M., and Beskow, J. (2012). "Animated speech: research progress and applications," in *Audiovisual Speech Processing*, eds R. B. Clark, J. P. Perrier, and E. Vatikiotis-Bateson (Cambridge: Cambridge University), 246–272.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Merabet, L. B., Rizzo, J. R., Amedi, A., Somers, D. C., and Pascual-Leone, A. (2005). What blindness can tell us about seeing again: merging neuroplasticity and neuroprostheses. *Nat. Rev. Neurosci.* 6, 71–77. doi: 10.1038/nrn1586

Miller, L. M., and d'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893. doi: 10.1523/JNEUROSCI.0896-05.2005

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., and Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clin. Linguist. Phon.* 20, 621–630. doi: 10.1080/02699200500266745

Moody-Antonio, S., Takayanagi, S., Masuda, A., Auer, J. E. T., Fisher, L., and Bernstein, L. E. (2005). Improved speech perception in adult congenitally deafened cochlear implant recipients. *Otol. Neurotol.* 26, 649–654. doi: 10.1097/01.mao.0000178124.13118.76

Nath, A. R., and Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J. Neurosci.* 31, 1704–1714. doi: 10.1523/JNEUROSCI.4853-10.2011

Nath, A. R., and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787. doi: 10.1016/j.neuroimage.2011.07.024

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* 95, 1085–1099. doi: 10.1121/1.408469

Niparko, J. K., Tobey, E. A., Thal, D. J., Eisenberg, L. S., Wang, N. Y., Quittner, A. L., et al. (2010). Spoken language development in children following cochlear implantation. *JAMA* 303, 1498–1506. doi: 10.1001/jama.2010.451

Nober, E. H. (1967). Vibrotactile sensitivity of deaf children to high intensity sound. *Laryngoscope* 78, 2128–2146. doi: 10.1288/00005537-196712000-00005

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cereb. Cortex* 17, 2251–2257. doi: 10.1093/cercor/bhl133

Osberger, M. J., Fisher, L., Zimmerman-Phillips, S., Geier, L., and Barker, M. J. (1998). Speech recognition performance of older children with cochlear implants. *Am. J. Otol.* 19, 152–157.

Ponton, C. W., Don, M., Eggermont, J. J., Waring, M. D., and Masuda, A. (1996). Maturation of human cortical auditory function: differences between normal-hearing children and children with cochlear implants. *Ear Hear.* 17, 430–437. doi: 10.1097/00003446-199610000-00009

Ponton, C. W., Moore, J. K., and Eggermont, J. J. (1999). Prolonged deafness limits auditory system developmental plasticity: evidence from an evoked potential study in children with cochlear implants. *Scand. J. Audiol.* 28(Suppl. 51), 13–22.

Ross, L. A., Saint-Amour, D., Leavitt, V. N., Javitt, D. C., and Foxe, J. J. (2007). Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

Rouger, J., Fraysse, B., Deguine, O., and Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Res.* 1188, 87–99. doi: 10.1016/j.brainres.2007.10.049

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., and Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7295–7300. doi: 10.1073/pnas.0609419104

Samuel, A. G., and Kraljic, T. (2009). Perceptual learning for speech. *Atten. Percept. Psychophys.* 71, 1207–1218. doi: 10.3758/APP.71.6.1207

Samuel, A. G., and Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 1479–1490. doi: 10.1037/a0036656

Schramm, D., Fitzpatrick, E., and Seguin, C. (2002). Cochlear implantation for adolescents and adults with prelinguistic deafness. *Otol. Neurotol.* 23, 698–703. doi: 10.1097/00129492-200209000-00016

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* 12, 106–113. doi: 10.1016/j.tics.2008.01.002

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Sharma, A., Dorman, M. F., and Spahr, A. J. (2002). A sensitive period for the development of the central auditory system in children with cochlear implants: implications for age of implantation. *Ear Hear.* 23, 532–539. doi: 10.1097/00003446-200212000-00004

Snik, A. F. M., Machdoum, M. J. A., Vermeulen, A. M., Brokx, J. P. L., and Van Den Broek, P. (1997). The relation between age at the time of cochlear implantation and long-term speech perception abilities in congenitally deaf subjects. *Int. J. Pediatr. Otorhinolaryngol.* 41, 121–131. doi: 10.1016/S0165-5876(97)00058-X

Song, J. J., Lee, H. J., Kang, H., Lee, D. S., Chang, S. O., and Oh, S. H. (2014). Effects of congruent and incongruent visual cues on speech perception and brain activity in cochlear implant users. *Brain Struct. Funct.* doi: 10.1007/s00429-013-0704-6. [Epub ahead of print].

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (London: Lawrence Erlbaum Associates, Inc.), 3–52.

Teoh, S. W., Pisoni, D. B., and Miyamoto, R. T. (2004). Cochlear implantation in adults with prelingual deafness. Part I. Clinical results. *Laryngoscope* 114, 1536–1540. doi: 10.1097/00005537-200409000-00006

Tjan, B. S., Chao, E., and Bernstein, L. E. (2014). A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. *Eur. J. Neurosci.* 39, 1323–1331. doi: 10.1111/ejn.12471

Trybus, R. J., and Karchmer, M. A. (1977). School achievement scores of hearing impaired children: national data on achievement status and growth patterns. *Am. Ann. Deaf* 122, 62–69.

Ungerleider, L. G., and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165. doi: 10.1016/0959-4388(94)90066-3

Waltzman, S., and Cohen, N. L. (1999). Implantation of patients with prelingual long-term deafness. *Ann. Otol. Rhinol. Laryngol. Suppl.* 177, 84–87.

Waltzman, S., Roland, J. T. Jr., and Cohen, N. L. (2002). Delayed implantation in congenitally deaf children and adults. *Otol. Neurotol.* 23, 333–340. doi: 10.1097/00129492-200205000-00018

Waltzman, S. B., Cohen, N. L., and Shapiro, W. H. (1992). Use of a multichannel cochlear implant in the congenitally and prelingually deaf population. *Laryngoscope* 102, 395–399. doi: 10.1288/00005537-199204000-00005

Wayne, R. V., and Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *J. Exp. Psychol. Appl.* 18, 419–435. doi: 10.1037/a0031042

Williams, P., and Simons, D. (2000). Detecting changes in novel, complex three-dimensional objects. *Vis. Cogn.* 7, 297–322. doi: 10.1080/135062800394829

Wilson, B. S., Dorman, M. F., Woldorff, M. G., and Tucci, D. L. (2011). Cochlear implants matching the prosthesis to the brain and facilitating desired plastic changes in brain function. *Prog. Brain Res.* 194, 117–129. doi: 10.1016/B978-0-444-53815-4.00012-1

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Commun.* 26, 23–43. doi: 10.1016/S0167-6393(98)00048-X

Yoon, Y. S., Li, Y., Kang, H. Y., and Fu, Q. J. (2011). The relationship between binaural benefit and difference in unilateral speech recognition performance for bilateral cochlear implant users. *Int. J. Audiol.* 50, 554–565. doi: 10.3109/14992027.2011.580785

Zeki, S. (2005). The Ferrier Lecture 1995: behind the seen: the functional specialization of the brain in space and time. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1145–1183. doi: 10.1098/rstb.2005.1666

Zeng, F.-G., Popper, A. N., and Fay, R. R. (2004). *Cochlear Implants: Auditory Prostheses and Electrical Hearing*. New York, NY: Springer.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 March 2014; paper pending published: 08 May 2014; accepted: 05 August 2014; published online: 26 August 2014.

*Citation: Bernstein LE, Eberhardt SP and Auer ET Jr. (2014) Audiovisual spoken word training can promote or impede auditory-only perceptual learning: prelingually deafened adults with late-acquired cochlear implants versus normal hearing adults. Front. Psychol. **5**:934. doi: 10.3389/fpsyg.2014.00934*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 May 2014
doi: 10.3389/fpsyg.2014.00366

Dynamic modulation of shared sensory and motor cortical rhythms mediates speech and non-speech discrimination performance

***Andrew L. Bowers*[1]*, Tim Saltuklaroglu*[2]*, Ashley Harkrider*[2]*, Matt Wilson*[3] *and Mary A. Toner*[1]**

[1] *Department of Communication Disorders, University of Arkansas, Fayetteville, AR, USA*
[2] *Department of Audiology and Speech Pathology, University of Tennessee Health Science Center, Knoxville, TN, USA*
[3] *School of Allied Health, Northern Illinois University, DeKalb, IL, USA*

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Nicolas Jean Bourguignon, Centre Hospitalier Universitaire Sainte Justine, Canada
Daniel Callan, Center for Information and Neural Networks, National Institute for Information and Communications Technology, Japan

***Correspondence:**
Andrew L. Bowers, Department of Communication Disorders, University of Arkansas, 606 North Razorback Road, Fayetteville, AR 72701, USA
e-mail: albowers@uark.edu

Oscillatory models of speech processing have proposed that rhythmic cortical oscillations in sensory and motor regions modulate speech sound processing from the bottom-up via phase reset at low frequencies (3–10 Hz) and from the top-down via the disinhibition of alpha/beta rhythms (8–30 Hz). To investigate how the proposed rhythms mediate perceptual performance, electroencephalographic (EEG) was recorded while participants passively listened to or actively identified speech and tone-sweeps in a two-force choice in noise discrimination task presented at high and low signal-to-noise ratios. EEG data were decomposed using independent component analysis and clustered across participants using principle component methods in EEGLAB. Left and right hemisphere sensorimotor and posterior temporal lobe clusters were identified. Alpha and beta suppression was associated with active tasks only in sensorimotor and temporal clusters. In posterior temporal clusters, increases in phase reset at low frequencies were driven by the quality of bottom-up acoustic information for speech and non-speech stimuli, whereas phase reset in sensorimotor clusters was associated with top-down active task demands. A comparison of correct discrimination trials to those identified at chance showed an earlier performance related effect for the left sensorimotor cluster relative to the left-temporal lobe cluster during the syllable discrimination task only. The right sensorimotor cluster was associated with performance related differences for tone–sweep stimuli only. Findings are consistent with internal model accounts suggesting that early efferent sensorimotor models transmitted along alpha and beta channels reflect a release from inhibition related to active attention to auditory discrimination. Results are discussed in the broader context of dynamic, oscillatory models of cognition proposing that top-down internally generated states interact with bottom-up sensory processing to enhance task performance.

Keywords: sensorimotor rhythms, independent component analysis, event-related spectral perturbations, intertrial coherence, speech perception

INTRODUCTION

A growing number of neurophysiological models have proposed that processes critical to receptive speech processing involve rhythmic cortical oscillations tuned to temporal regularities of speech (Arnal and Giraud, 2012; Giraud and Poeppel, 2012; Peelle and Davis, 2012; Ghitza, 2013). On the production side, theories (e.g., frame/content theory) propose that the auditory system has been tuned to the quasi-periodic constraints imposed by articulator movements (MacNeilage, 1998). On the receptive side, oscillatory frameworks posit that the articulatory-motor system structures its output to match rhythms best captured by the auditory system at multiple timescales (Giraud and Poeppel, 2012; Peelle and Davis, 2012). A fundamental link between the speech production mechanism giving rise to the acoustic signal and rhythmic sampling of the same signal in sensorimotor networks would be advantageous for a neural system tasked with resolving highly variable acoustic cues (Callan et al., 2010). However, it is as yet unknown how rhythmic processes in motor and sensory regions are integrated on a millisecond time scale and it remains unclear under what conditions sensorimotor integration is adapted to improve perceptual outcomes (Gallese et al., 2011).

According to internal model theories of speech production, neural connections between perception and production are tuned as infants learn to produce auditory targets (Callan et al., 2000). Neurophysiological dual-stream models suggest that this auditory to articulatory link is accomplished via a network of regions known as the dorsal stream, including primary auditory and auditory association areas, inferior parietal regions, and areas of the premotor and sensorimotor cortex (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Specht, 2014). In accordance with this proposal, models of cortical rhythm generation have suggested that neural oscillations in overlapping frequency bands in auditory regions are tuned over time to the natural rhythms of speech production (Poeppel, 2003; Giraud et al., 2007; Morillon et al., 2010; Giraud and Poeppel, 2012). More specifically, over the course of development the motor and premotor

cortex may tune the response of sensory regions to natural low-frequency rhythms associated with jaw and lip movements corresponding to syllabic rate (∼4 Hz). Thus, motor regions involved in initiating speech movements and auditory areas involved in parsing speech are thought to share common temporal framework.

On the receptive side, it has long been suggested that the syllable unit may represent an integrative time window in which phonemes occurring at higher rates (∼20–50 ms) are processed as part of a longer temporal unit (∼100–250 ms) occurring at slower rates (Massaro, 1974). In support of that notion, psychophysical data suggest that categorization of speech stimuli along a continuum occurs in an integration window between 110 and 150 ms consistent with one cycle of oscillation in the theta band (Chang et al., 2010). Higher-order auditory association areas (e.g., BA 22) also show the property of theta–gamma nesting in which the phase of fast gamma rhythms (50–70 Hz) is locked to the phase of slower theta rhythms (4–8 Hz), suggesting that phonemic categorization is integrated in a time window consistent with the syllable unit (Giraud and Poeppel, 2012). Low-frequency rhythms (∼3–10 Hz) have also been implicated more generally in sensorimotor integration (Bland and Oddie, 2001), indicating they may provide a common mechanism by which sensory and motor systems share information for a range of sensory signals associated with previous sensorimotor experience (Poeppel, 2003; Giraud et al., 2007; Morillon et al., 2010; Giraud and Poeppel, 2012). As such, during receptive speech processing, it has been proposed that sensory and motor oscillatory assemblies tuned to the expected temporal structure of speech reset to aid in the organization of continuous, stimulus driven neural spike trains into abstract units for further analysis (Giraud and Poeppel, 2012). Importantly, while delta–theta phase reset consistent with the syllable unit has been demonstrated in a studies using continuous, phrase and sentence level auditory stimuli (Luo and Poeppel, 2007; Doelling et al., 2013), it has not been demonstrated to play a role in sensorimotor integration during speech perception.

In addition to the potential role of low-frequency rhythms in sensorimotor integration, recent theoretical frameworks have implicated a functional role for beta rhythms both in motor control and perception (13–30 Hz; Engel and Fries, 2010; Arnal and Giraud, 2012). On the motor side, beta band activity is associated with the rolandic sensorimotor rhythm. The sensorimotor rhythm is thought to reflect processing downstream from premotor regions and is associated with source estimates clustering near the central sulcus. In particular, suppression of the beta band (∼20 Hz) over the sensorimotor cortex is associated with the observation, imagination, and execution of movements in a somatotopic manner (see Hari, 2006 for review). It has been proposed that efferent copies of a motor goal transmitted along beta channels suppress responses in sensory regions to the expected event, freeing the sensory system to respond to external sensory stimuli (Engel and Fries, 2010). Efferent copies of expected sensory events have also been shown to have a significant effect on the interpretation of upcoming sensory cues (Driver and Frith, 2000; Frith et al., 2000). As such, a common function for beta band oscillations in both motor control and perceptual contexts may be to generate top-down influences functioning to override unexpected sensory events or conversely to enhance activity focused on expected sensory features (Engel and Fries, 2010).

More recently, Arnal and Giraud (2012) proposed that beta rhythms interact with low frequency rhythms to enhance processing focused on anticipated sensory events. According to their framework, attention to behaviorally relevant task goals and temporal predictions about expected sensory events modulate oscillatory processing in two ways. When sensory events can be predicted in time (e.g., in connected speech) delta–theta oscillations reset prior to stimulus onset in anticipation of the forthcoming event, reflecting a predicting "when" scheme. However, when a sensory event cannot be predicted in time, delta–theta phase reset is commensurate with stimulus onset. In that case, a predicting "what" scheme may apply in which top-down, content related sensory predictions transmitted along beta channels interact with low-frequency phase reset during the sensory event. The functional effect of the two complementary mechanisms is to boost the gain of neural responses to sensory signals within the attended or temporal focus. Given the proposed role of beta rhythms in both sensory prediction and motor control, the sensorimotor cortex appears to be in a good position to process incoming information from the bottom-up at delta–theta frequencies and to be involved in top-down content related predictions at beta frequencies (i.e., prior to the event).

Along with low-frequency rhythms and beta band activity, some recent accounts have also emphasized the potential role of alpha (8–12 Hz) rhythms in speech processing. Obleser et al. (2012) suggested that disinhibition along alpha channels may function to enhance sensory processing to attended auditory events. According to general models of alpha function, high power in the ongoing alpha band is viewed as an active inhibitory mechanism functioning to gate irrelevant information, permitting increased processing focused on events relevant to task goals when disinhibited (Klimesch et al., 1998). In addition, alpha disinhibition (i.e., decrease in band power) prior to sensory input has been shown to predict accurate task performance in visual perception tasks, suggesting a top-down modulatory role in perceptual performance (Fellinger et al., 2011). In accordance with these proposals, a growing body of evidence implicates an auditory alpha generator in the temporal lobes that may suppress during active attention to auditory stimuli. Suppression within the traditional alpha band has been recorded near the auditory cortex and auditory association areas using electrocorticographic (ECoG) recordings (Crone et al., 2001). Alpha suppression localized to the primary and auditory association areas has also been demonstrated during auditory attention to contralateral acoustic stimuli and noise-vocoded word comprehension prior to and following the auditory signal (Weisz et al., 2011; Obleser et al., 2012). Scalp-recorded electroencephalographic (EEG) recordings have shown that upper alpha rhythms (10–12 Hz) suppress during effortful, sublexical speech processing (Cuellar et al., 2012). Thus, much like beta suppression, alpha disinhibition may play a functional role in auditory attention, functioning to facilitate processing focused on expected sensory events (Callan et al., 2010; Weisz et al., 2011).

Although it is still unclear how local neuronal assemblies in the dorsal stream share information globally, Giraud and Poeppel (2012) propose that motor and sensory regions process information across a broad spectral range consistent with processing at multiple time scales. According to that model, an intrinsically left hemisphere dominant region in the lip and tongue area motor cortex (1–72 Hz) and hand area motor cortex (2–6 Hz) is connected to the somatosensory and auditory regions (1–72 Hz). The proposed functional role of input from the lip area is to contribute to the parsing of speech at syllabic rates in sensory regions, suggesting that the lip region is involved in the modulation of low frequency rhythms (Morillon et al., 2010; Giraud and Poeppel, 2012). While the model does not specify a role for the motor system in sensory prediction, current internal model frameworks suggest that early sensorimotor models in the same region may function to constrain sensory analysis when sensory cues are ambiguous (Skipper et al., 2006; Callan et al., 2010) or to boost the gain of assemblies tuned to expected sensory features similarly to the proposed role of selective attention in visual perception (Hickok et al., 2011). Given proposals that auditory and sensorimotor regions are tuned over the course of development along shared oscillatory channels, it is reasonable to hypothesize that the two regions would show activity in the same oscillatory bands relevant to sensory processing at delta–theta, alpha, and beta frequencies.

In light of oscillatory frameworks, potential differences in activity along shared oscillatory bands within locally synchronized regions would be predicted to vary depending on internal states of expectancy and bottom-up sensory input. The available neuroimaging evidence in sublexical speech discrimination tasks supports the notion that dorsal stream sensorimotor activation does indeed vary with internal state, task goals, and bottom-up input (Binder et al., 2004; Callan et al., 2010; Osnes et al., 2011). In a passive task, Osnes et al. (2011) demonstrated that, as white noise is parametrically morphed into the acoustic structure of speech syllables, an area within the dorsal premotor cortex is active only at an intermediate step related to perceptual ambiguity. That finding suggests that when attention is not directly allocated to phoneme discrimination, the premotor cortex is only active when acoustic cues are ambiguous. However, another study using the same sound morphing procedure also demonstrated that when participants told to expect vowels or musical notes prior to stimulus presentation, ventral and dorsal aspects of the premotor cortex extending into sensorimotor regions were active, suggesting that top-down anticipatory processes are associated with motor activation even in the absence of a task (Osnes et al., 2012). Other studies using passive tasks have also reported activity in the premotor and somatomotor areas when participants listened to trains of repeated syllables (Wilson et al., 2004; Pulvermüller et al., 2006). Importantly, while activity in motor regions clearly occurs in passive tasks, a wide range of explanations have been proposed to explain why it occurs and how it functions. It has been suggested that motor activity may be related to resolving perceptual ambiguity (Osnes et al., 2011), in some cases covert rehearsal of repeated syllable trains (Hickok et al., 2011), or more recently may be modulated by states of expectancy even in the absence of task goals (Osnes et al., 2012). However, as perceptual performance cannot be assessed in passive listening it is unclear in such conditions whether motor activity plays a functional role in perceptual performance (Adank, 2012).

Whereas performance related brain activity cannot be assessed in passive tasks, it may be investigated using active tasks in which participants register a response. Using a two-forced choice discrimination task in which attention was directed to phoneme discrimination, Binder et al. (2004) demonstrated that blood-oxygen level dependent (BOLD) signals in auditory association areas decrease as background noise increases, suggesting that bottom-up acoustic cues are critical to activation of auditory regions. However, as auditory signal degradation increased, greater activity was observed in the posterior portion of Broca's area, suggesting that premotor regions play a compensatory role when acoustic cues are degraded. Employing a similar experimental paradigm, Callan et al. (2010) demonstrated that BOLD and constrained time–frequency measures using MEG in posterior temporal lobe regions were not associated perceptual performance in in noise (i.e., correct relative to incorrect trials). However, dorsal and more ventral regions of the left hemisphere premotor system were related to perceptual performance. MEG analysis indicated alpha and beta suppression both prior to and following correct discrimination trials in the more ventral region of the PMC. The findings were interpreted within internal model frameworks suggesting that efferent articulatory models initiated in motor regions function to constrain sensory analysis in noisy listening conditions and mediate perceptual performance. Consistent with that study, Alho et al. (2012) demonstrated an early (∼100 ms) potential following stimulus input in a region of interest within the precentral gyrus that was greater for passive relative to active discrimination in noise, supporting proposed explanations for observed differences in motor activity during passive relative to active tasks.

Studies using transracial magnetic stimulation (TMS) during active speech discrimination/identification tasks have largely corroborated functional imaging findings. Stimulation to dorsal stream premotor regions results in increased reaction times following tasks requiring speech segmentation without noise (Sato et al., 2009) and enhanced adaptation to speech stimuli (Grabski et al., 2013). TMS stimulation to the primary motor cortex (M1) during active tasks has also been shown to facilitate speech identification for the effector involved, suggesting an effector specific function in perceptual constraints at the level of anticipated spectro-temporal features (D'Ausilio et al., 2009). In addition, studies using passive tasks have shown impaired categorical perception (Möttönen and Watkins, 2009) and reduced auditory event-related potentials (ERPs) with stimulation to the lip region of M1 (Möttönen et al., 2013), suggesting that stimulation to motor regions may modulate auditory processing even in the absence of a task. Although it is unclear why some studies have shown more ventral activity along the precentral gyrus and others have shown activity or performance related differences with stimulation in the more dorsal premotor cortex and adjacent somatomotor regions (Möttönen and Watkins, 2012), it is clear that both regions play some role in speech perception tasks.

Further insight into what drives differences in regional activation across internal states, task goals, and levels of bottom-up input might be derived from a "dynamicist" model of cognition (Engel

et al., 2001; Fries, 2005; Siegel et al., 2012; and see Callan et al., 2010 for application to speech perception). According to the this view, top-down influences may be defined as endogenously generated sources of contextual modulation supporting large-scale thalamocortical and cortico-cortical interactions in goal-definition, action planning, working memory, and selective attention (Engel et al., 2001; Fries, 2005). Neural synchrony in the millisecond range is taken to be critical for processing incoming sensory signals, not only in higher order sensory association areas, but as a result of synchrony between regions involved in previous experience, including the procedural knowledge stored in sensorimotor networks (Driver and Frith, 2000; Frith et al., 2000). According to this model, during sensory perception top-down influences carry predictions about feature constellations that are then matched with bottom-up sensory input in a manner similar to analysis-by-synthesis (Stevens and Halle, 1967; Poeppel and Monahan, 2011). These shared modulatory influences are also thought to compete for stable resonant states reflecting a best match between internal states of expectancy and bottom-up sensory features. As such, degenerate neural mappings between regions may function flexibly in different oscillatory patterns depending on internal states, perceptual context, and bottom-up sensory cues to achieve the same perpetual outcomes (i.e., invariant categorization; Engel et al., 2001).

From a dynamic perspective, active and passive perceptual tasks may involve different mechanisms within the same sensorimotor network. In passive tasks, bottom-up sensory cues processed in temporal lobe regions may drive enhanced activity in motor regions when spectro-temporal cues are ambiguous. In that case, enhanced activity in motor regions might function to produce a more stable match between ambiguous sensory cues and corresponding representations in the motor cortex (Osnes et al., 2011) or in some cases to aid working memory for more complex tasks (Sato et al., 2009). However, when attention is directed to a discrimination task, motor regions closely linked to expected sensory features may function to monitor internal states related to attention task goals, with greater activity in motor or auditory regions when bottom-up sensory cues match with expected sensory features. In that case, the sensorimotor cortex could be characterized as one component of an entrained network involved in phonological or articulatory selective attention prior to and throughout sensory processing (Skipper et al., 2006; Callan et al., 2010; Hickok et al., 2011). From a dynamacist viewpoint, whether higher order auditory regions or sensorimotor regions are selectively enhanced would depend on which local region provides the best match between predicted feature constellations and bottom-up input (Engel et al., 2001; Fries, 2005). Thus, if the sensorimotor cortex plays a specific role in articulatory selective attention, early sensorimotor activity prior to stimulus presentation with subsequent response amplification following sensory input would only be expected for acoustic stimuli closely associated with articulatory production (i.e., syllables). Further, speech specific enhancement would be expected to occur only when bottom-up spectro-temporal cues are sufficient to support successful discrimination.

Initial evidence consistent with a role for the motor system in articulatory selective attention was reported in a recent study (Bowers et al., 2013). In that study, to address the role of the sensorimotor cortex in passive and active contexts, event-related EEG was used to measure oscillatory activity of the rolandic sensorimotor μ rhythm prior to, during, and following a speech and non-speech discrimination task in varying levels of white noise. A blind source separation approach (BSS) known as independent component analysis (ICA) was used to isolate the sensorimotor rhythm from other volume-conducted components of the EEG signal (Delorme et al., 2012). Although no changes in power relative to baseline were observed in passive tasks, early left-hemisphere beta (15–25 Hz) suppression localized to the lateral central sulcus was observed prior to stimulus onset, with peak suppression just following auditory stimuli for syllables only. Peak suppression just following acoustic events for correct trials in high SNR (+4 dB) conditions was also greater than for the same syllable discrimination task at a low SNR (–6 dB) in which participants performed at chance. Due to the time-course of beta activation and speech selective responses, the findings could not be attributed to covert rehearsal or simple sensory-decision mechanisms (Bowers et al., 2013). Early sensorimotor beta suppression prior to stimulus onset was interpreted as an articulatory model functioning to constrain sensory analysis, with decreases in activity when initial hypotheses were at odds with bottom-up input. However, as the analysis was confined to the sensorimotor rhythm and a measure of power only, it was unclear how bilateral posterior auditory components also submitted by ICA functioned in those tasks. Given the predictions of current oscillatory frameworks, the sensorimotor and auditory association regions would be expected to share cortical rhythms at delta–theta, alpha, and beta frequencies varying as a function of task and bottom-up sensory input.

To address how proposed sensory and sensorimotor rhythms function in the performance of a speech and non-speech discrimination task, the aims of the current analysis are: (1) to investigate whether alpha-like posterior temporal lobe clusters are also associated spectral suppression along shared at beta and alpha frequencies surrounding and during stimulus events; and (2) to investigate how cortical rhythms shared between sensorimotor and temporal lobe clusters vary depending on task and the quality of bottom-up acoustic input (i.e., correct relative to chance trials). Within the context of current frameworks, a number of predictions can be made about how cortical rhythms vary in time, frequency, and space during passive listening and an anticipatory speech and non-speech discrimination task. First, ICA is expected to reveal an independent alpha-like generator with scalp-topographies over the posterior temporal lobes and source estimates in auditory association areas. Second, consistent with previous findings, alpha suppression in posterior temporal lobe regions would be expected prior to, during, following auditory stimuli in active tasks in which attention is directed to discrimination. However, if as current oscillatory frameworks posit (e.g., analysis by synthesis), sensorimotor regions associated with speech articulation participate in top-down predictions along beta channels, auditory regions would be expected to suppress along in the same oscillatory band during active tasks. Third, if low frequency phase reset (3–10 Hz) is associated with bottom-up mechanisms only, it would be expected in auditory regions regardless of the

task or type of stimulus input, with a decrease when bottom-up sensory cues are insufficient for discrimination (i.e., chance trials) relative to trials in which spectro-temporal cues are clear (i.e., correct trials). However, given the predictions of dynamic oscillatory frameworks, another possibility is that performance related selective responses along delta–theta channels compete during sensory input, reflecting the influence of both top-down and bottom-up mechanisms. In that case, during active processing, if the sensorimotor cortex plays a specific role in articulatory selective attention, it would be expected to increase in active tasks generally with further enhancement when bottom-up sensory input matches with expected sensory features (i.e., correct trials) and to decrease when such expectations were not fulfilled (i.e., chance trials). Further, a pattern consistent with efferent motor models would be expected for speech stimuli but not tone-sweep stimuli.

MATERIALS AND METHODS

PARTICIPANTS

Sixteen right-handed English-speaking adults (15 female and 1 male) with a mean age of 25 (range 20–42) participated in this study. Participants were recruited from the general population at the University of Tennessee. Participants reported no diagnosed history of communicative, cognitive, or attentional disorders. Degree of handedness was assessed using the Edinburg Handedness inventory (Oldfield, 1971). This study was approved by the Institutional Review Board of the University of Tennessee Health Science Center. Prior to the experiment, all participants were provided with an informed consent document approved by the Institutional Review Board and all participants gave written informed consent prior to inclusion.

STIMULI

Speech stimuli consisted of /ba/ and /da/ syllable generated using AT&T naturally speaking text-to-speech software. The software generates syllables from text using speech synthesized from a human male speaker. Half of the stimuli were composed of different initial sounds (e.g., /ba/ and /da/) and the other half were the same (e.g., /ba/ and /ba/). The stimuli were normalized to have the same root-mean-square (RMS) amplitude and low-pass filtered with a cutoff at 5 kHz. Each stimulus syllable was 200 ms in duration with an interstimulus interval of equal length (i.e., 200 ms). Thus, the total time required to present a stimulus pair was 600 ms. For the tone discrimination task, sine-wave tone sweeps were generated using a procedure adapted from a previous neuroimaging study (Joanisse and Gati, 2003). Tone-sweep stimuli were composed with an 80 ms modulated tone onset and a 120 ms steady state 1000 Hz sine-wave.

As with the speech stimuli, tone-sweeps were generated, low-pass filtered with a cut-off at 5 kHz, and normalized to have the same RMS amplitude as the speech stimuli. Tone pairs differed only in whether the pitch onset was lower at 750 Hz than the steady state tone or higher at 1250 Hz. For both speech and tones the intertrial interval was 3000 ms. White noise for the tone and speech stimuli was generated and processed using the same procedure as for the speech sounds, with a low-pass filter cut-off at 5 kHz. All auditory stimuli were processed using Soundtrack Pro academic software on an iMac (2 GHz Intel core duo) computer and were sampled at 44 kHz. Conditions were placed in random order prior to presentation. All stimuli were presented at an absolute intensity of ∼70 dB.

Previous investigations have shown better than chance performance on a forced choice syllable discrimination task using a +4 dB SNR and chance performance using a –6 dB SNR (Binder et al., 2004; Callan et al., 2010). However, pure tones may be detected with noise intensities as high as 18 dB above pure tone intensity (i.e., –18 dB SNR; Ernst et al., 2008). To account for differences in perceived loudness between tone and speech stimuli, preliminary behavioral data were collected from 10 participants using Stim2 presentation software presented through Etyomotic ER1-14A tube phone inserts in a sound-treated booth. Syllable and tone stimuli were embedded in white noise and presented in 20 trials at the following SNRs –18, –12, –6, +4 dB. Syllable stimuli were identified above chance in the +4 dB condition only. Accuracy for tone-sweep conditions were not above chance in -18 dB SNR, with 60% in –12 dB SNR, 78% in the –6 dB condition, and 76% in +4 dB condition. Paired *t*-tests revealed no significant difference ($p > 0.05$) between the +4 and –6 dB tone-sweep conditions. As such, the SNRs for the syllables were set at +4 and –6 dB and for tone-sweeps at +4 and –18 dB.

PROCEDURE

Stimuli were presented using Stim 2 4.3.3 stimulus presentation software on a PC computer. The experiment was conducted in an electronically and magnetically shielded, double-walled, sound-treated booth. Participants were seated in a comfortable reclining armchair with their heads and necks well supported. Participants were told that they would be listening to white noise, syllables, and tones. They were instructed that the onset of one trial would commence when white noise was audible, followed by either syllable or tone stimuli. Participants were asked to indicate whether the syllables or tone-sweeps sounded the same or different by pressing a button using the left thumb only. To further control for the possibility that preparation for the response might confound motor activity related to stimulus processing, participants were signaled to respond via a 100 ms, 1000 Hz sine wave tone 1400 ms after stimulus onset. To control for stimulus–response bias in the button press task, the order of the button press was counterbalanced (Callan et al., 2010).

All conditions were randomized prior to presentation and presented in two randomized blocks consisting of 40 trials each. Performance was evaluated as a percentage of correct trials (%CT) and response time (RT). Participants were asked to listen under the following conditions: (1) passively listening to noise (PasN); (2) passively listening to speech syllables in +4 dB noise (PasSp + 4 dB); (3) passively listening to tone-sweeps in +4 dB noise (PasTn + 4 dB); (4) active syllable discrimination-in +4 dB noise (ActSp + 4 dB); (5) active tone-sweep discrimination-in +4 dB noise (ActTn + 4 dB); (6) active syllable discrimination in –6 dB noise (ActSp – 6 dB); (7) active tone-sweep discrimination in –18 dB noise (ActTn – 18 dB).

EEG ACQUISITION

Thirty-two channels were used to acquire EEG data based on the extended international 10–20 method of electrode placement

using an unlinked, sintered NeuroScan Quik Cap (Jasper, 1958). Recording electrodes included FP1, FP2, F7, F3, FZ, F4, F8, FT7, FC3, FCZ, FC4, FT8, T7, C3, CZ, C4, T8, TP7, CP3, CPZ, CP4, TP8, P7, P3, PZ, P4, P8, O1, OZ, O2 with two electrodes on the left (M1) and right mastoids (M2). The reference electrode was placed on the nasion and the ground electrode was at FPZ. The electro-oculogram (EOG) was recorded by electrodes placed on the left superior orbit and the left inferior orbit (VEOG) and on the lateral and medial canthi of the left eye (HEOG) to monitor vertical and horizontal eye movements, respectively. The impedances of all electrodes were measured at 30 Hz before, during, and after testing and were never greater than 5 kΩ.

EEG data were collected using Compumedics NeuroScan Scan 4.3.3 software and the Synamps 2 system. The raw EEG data was filtered (0.15–100 Hz), and digitized via a 24-bit analog-to-digital converter at a sampling rate of 500 Hz. Data was time-locked to the onset of individual speech perception trials. After data collection, the recorded EEG signal and EOG data was segmented into single trials lasting approximately 5000 ms each, spanning from –3000 to +2000 ms with reference to stimulus onset (i.e., zero time). To examine pre- and post-stimulus activity, the EEG data were epoched into 5000 ms segments. EEG data were visually inspected and trials contaminated by gross artifacts greater than 200 μV were removed. A minimum contribution of 40 epochs for each participant in each condition was required for inclusion in the experiment. Due to a contribution of only 20 trials in several conditions, one participant was omitted from analysis.

ICA PREPROCESSING

To decrease computational requirements for ICA processing, data were downsampled to 256 Hz. Prior to ICA training, EEG data were concatenated for each participant across conditions. Subsequent ICA training was implemented using the extended runica algorithm implemented in EEGLABv12. The initial learning rate was set to 0.001 with a stopping weight of 10–7. Linear decomposition using the extended Infomax algorithm (Lee et al., 1999) was conducted for each participant across experimental conditions. The algorithm spheres the data matrix prior to ICA rotation and computes the variance of IC projection weights on to the original EEG channel data (Delorme and Makeig, 2004). The resulting square weight matrix (30×30) is thus applied to each participant, yielding a single set of weights for each experimental condition expressing independence in the data. The inverse weight matrix (W^{-1}) can then be projected onto the original EEG channel configuration, providing a spatial scalp topography for the components.

Independent components (ICs) were evaluated for each participant across experimental conditions using three criteria. First, an automated algorithm (ADJUST) shown in a previous study to have good inter-rater reliability with researchers experienced in IC noise removal, was used to tag non-brain artifact components in the EEGLAB module (Mognon et al., 2010). Scalp-maps and log spectra were also visually inspected for indicators of non-brain artifact including abnormal spectral slope, and scalp-topographic distributions known to be associated with eye-movement and temporal muscle contraction (Onton and Makeig, 2006). Second, ICs with 20 trials having outlier values (μV SD set to 10) over the electrode with maximum power were eliminated (Callan et al., 2010). Finally, equivalent current dipole (ECD) models for each component were computed using a standard template boundary element model (BEM) in the DIPFIT toolbox, freely available at sccn.ucsd.edu/eeglab/dipfit.html (Oostenveld and Oostendorp, 2002). As individual magnetic resonance (MR) structural models were not available, 10–20 electrode coordinates assuming a common head shape were warped to the standard template head model followed by automated coarse and fine-fitting, yielding dipole models for each of 480 ICs. The procedure involves hypothesizing a dipole source that could have generated the scalp potential distribution for a given IC and then computing the model that explains the highest percentage of the variance in the scalp map (Delorme et al., 2012).

sLORETA SOURCE ESTIMATIONS

sLORETA is a functional imaging technique that provides standardized linear solutions for modeling 3D distributions of the likely cortical generators of EEG activity (Pascual-Marqui, 2002). The software uses a 3D spherical head model separated into compartments including, the scalp, skull, and brain. sLORETA analysis operates under the assumption that scalp-recorded signals originate primarily in the cortical gray matter/hippocampi and that neighboring neurons are synchronously activated, giving rise to a signal that is distinct from surrounding noise. The head model is standardized with respect to the Talairach cortical probability brain atlas, digitized at the Montreal Neurological Institute (MNI) and uses EEG electrode coordinates derived from cross-registrations between spherical and realistic head geometry (Towle et al., 1993). The brain compartment includes 6239 voxels (5 mm resolution). Electrode coordinates were exported to sLORETA from the EEGLAB module. For each IC, inverse ICA weight projections onto the original EEG channels were exported to the sLORETA data processing module for each participant. Cross-spectra were computed and mapped to the standard Taliarach brain atlas cross-registered with the MNI coordinates, yielding sLORETA estimates of current source density (CSD) for each of 480 ICs.

INDEPENDENT COMPONENT CLUSTERING

To identify similar ICs across participants, 480 (30×16) components were then clustered using measure product methods in the K-means toolbox implemented in EEGLAB (Delorme and Makeig, 2004). The toolbox uses principle component clustering methods to reduce data dimensions and yields similar component clusters across participants. Here, 28 possible component clusters were considered. The data dimensions were reduced to 10 with the standard deviation set to 3. As such, ICs more than 3 SDs from any cluster mean were excluded as an outlying cluster. As both the auditory alpha and sensorimotor components are thought to have distinct spectral signatures, scalp-topographies, and source estimates were precomputed and used in the clustering analysis. Component power spectra for each subject were calculated by averaging fast Fourier transform (FFT) spectra for each epoch using a window length of 256 points. Scalp topographies were computed as 30 channel (x,y) map gradients. ECD models and sLORETA CSD distributions for each participant were precomputed in the

manner described in a previous section. Only components with a single dipole model within the head volume accounting for 80% or greater of the variance in the IC scalp distribution were included in component clusters. Pre-identified noise components tagged prior to the analysis were used to identify clusters accounting for non-brain sources. Given the initial hypotheses of a posterior temporal lobe alpha rhythm and well-known spectral signatures for the sensorimotor rhythm (see Bowers et al., 2013), only components with distinct spectral peaks near 10 Hz for components with a temporal distribution and those with peaks at ∼10 and ∼20 Hz for those with a sensorimotor distribution were included in temporal and sensorimotor clusters, respectively.

To examine stimulus induced changes in the EEG, time–frequency transforms were precomputed in the EEGLAB module using the STUDY command structure. A measure of power (event-related spectral perturbations ERSPs) and a measure of phase (intertrial coherence ITCs) were used to investigate ICA activation. ERSPs are changes scaled in normalized decibel units over a broad spectral range (here 3–40 Hz) and ITCs are a measure of the strength of phase alignment across trials (Delorme and Makeig, 2004) and have been used to measure stimulus phase alignment in previous studies of sentence level speech processing (e.g., Luo and Poeppel, 2007). For ICs, ERSPs are scaled in RMS decibel units on the same scale as the component and ITCs are represented via a magnitude scale from 0 (weakest) to 1 (strongest). In this study, time–frequency transforms were computed using a Morlet sinusoidal wavelet set at three cycles at 3 Hz rising linearly to 20 cycles at 40 Hz. A 1000 ms pre-stimulus baseline was selected from the silent intertrial interval. This baseline served as a time period during which a surrogate distribution was generated. The surrogate data distribution is constructed by selecting spectral estimates for each trial from randomly selected latency windows in the specified epoch baseline. In this study, the baseline data was sampled 200 times, producing a baseline distribution whose percentiles were taken as significance thresholds (Makeig et al., 2004). Significant changes in ERSPs or ITC magnitude (i.e., increases or decreases from the silent recording interval) were then tested using a bootstrap resampling method. Significant differences from baseline ($p < 0.05$ uncorrected) were considered in the subsequent within subjects analysis of both ERSPs and ITCs.

Analysis of condition effects was carried out using the STUDY command structure in EEGLAB. The single trial current for all seven experimental conditions for frequencies between 3 and 40 Hz and times from –600 to 1500 ms post-stimulus onset were entered into a time–frequency analysis. For the two conditions in which performance was better than chance (ActSp + 4 dB and ActTn + 4 dB) only trials discriminated correctly were considered in the ERSP analysis. A mean of 64 trials across conditions were entered into the ERSP and ITC analysis. Wavelet estimates across trials for each time and frequency were then converted to a time–frequency matrix (69 × 105) from 3.4 to 39.9 Hz to –589 to 1441 ms. To test the significance of condition effects, non-parametric random permutation statistics adopting a 1 × 7 repeated measures ANOVA design were computed. The random distribution represents the null hypothesis that no condition differences exist. In the current study, 2000 random permutations were computed and compared to *F*-values for the mean condition differences. To control for the inflation of type I error rates associated with multiple comparisons, a correction for false discovery rate (*p*FDR) was applied, allowing for a conservative test of condition effects (Benjamini and Hochberg, 1995).

RESULTS

PERCENTAGE CORRECT TRIALS

Prior to the analysis, trials with RTs greater than three standard deviations from the mean RT (i.e., trials greater than 1996 ms) were removed and were not considered in any subsequent analysis. Performance on the active perceptual identification tasks (i.e., tasks in which a response was required) was assessed as a percentage of correct trials. However, as it has been demonstrated that premotor and sensorimotor regions are sensitive to response bias in a speech discrimination task as opposed to perceptual sensitivity (Venezia et al., 2012), *d'*-values are also reported. For the active conditions, a repeated measures analysis of variance (ANOVA) with the factor condition (1 × 4) revealed a significant main effect [$F_{(3,45)} = 131.65, p = 0.00$]. A series of paired comparisons with a Bonferroni correction for the number of comparisons was employed to determine condition differences. A comparison between ActSp + 4 dB and ActSp – 6 dB [$F_{(1,15)} = 207, p = 0.000, \eta^2 = 0.96$] and between ActTn + 4 dB and ActTN – 18 dB [$F_{(1,15)} = 113, p = 0.00, \eta^2 = 0.88$] indicated greater %CT in the two high SNR conditions. A significant difference was found for a comparison between %CT in the ActSp + 4 dB condition and the ActTn + 4 dB condition [($F_{(1,15)} = 39, p = 0.00, \eta^2 = 0.72, \Phi = 1$]. No significant difference was found for a comparison of the Actsp – 6 dB and Actn – 18 dB conditions [$F_{(1,15)} = 1.79, p = 0.20$]. The ActSp – 6 dB and ActTn – 18 dB were also not significantly different from chance ($t = 0.98, p = 0.20$). Thus, performance in the ActSp + 4 dB condition [96% (SE = 0.01); $d' = 3.25$ (SE = 0.14)], was higher than performance in the ActTn + 4 dB condition [83% (SE = 0.02); $d' = 1.61$ (SE = 0.22)]. The means for the ActSp – 6 dB and ActTn – 18 dB were not significantly greater than chance at [52% (SE = 0.01); $d' = 0.13$ (SE = 0.11)] and [51% (SE = 0.01); $d' = 0.07$ (SE = 0.11)], respectively. Thus, as expected, only the speech and tone-sweep conditions with a relatively high SNR were associated with better than chance performance.

RESPONSE TIME

RTs for each subject in the four active conditions were entered into a repeated measures ANOVA with the factor condition (1 × 4). The analysis revealed a significant main effect for condition [$F_{(3,45)} = 3.71, p = 0.010, \eta^2 = 0.19, \Phi = 0.77$]. Planned comparisons with Bonferroni adjustments revealed no significant difference between the Actsp + 4 dB and ActTn + 4 dB conditions [$F_{(1,15)} = 0.00\ p = 0.96$] or between the ActSp − 6 dB and ActTn − 18 dB [$F_{(1,15)} = 0.24\ p = 0.62$]. A comparison of correct trials in the ActSp + 4 dB and ActTn + 4 dB compared to chance trials in the ActSp − 6 dB and ActTn − 18 dB conditions, respectively, revealed a significant difference [$F_{(1,15)} = 7.23, p = 0.016, \eta^2 = 0.32, \Phi = 0.71$], indicating that correct trials were associated with a lower mean RT than chance trials. The mean RT for the two conditions in which performance (ActSp + 4 dB

and ActTn + 4 dB) was above chance were 642 ms (SE = 58) and 641 ms (SE = 47), respectively. The mean RT for the two conditions in which performance was at chance levels was 767 (SE = 68) and 743 ms (SE = 55), respectively. Taken together, the analysis of behavioral responses revealed an inverse relationship between perceptual performance in the active conditions and button press RT.

INDEPENDENT COMPONENT CLUSTERING

Independent component clustering revealed eight distinct component clusters with neural as opposed to non-brain (i.e., artifact) sources. Six component clusters accounted for eye-blinks, vertical eye-movements, horizontal eye-movements, temporal muscle noise, and non-specific noise (electromagnetic noise). Component clusters with similar scalp-topographies, spectra, ECD, and sLORETA CSD locations were found for a left hemisphere frontal, frontal midline cluster, central midline cluster, left and right sensorimotor clusters, and left and right posterior temporal clusters. A less consistent (10 ICs) left-hemisphere parietal cluster was also identified. However, as the focus of the current investigation is on the sensorimotor and posterior temporal clusters, only these clusters are discussed further.

For the posterior temporal clusters, thirteen participants submitted ICs with topographic distributions over the left temporal lobes and thirteen participants submitted ICs with right hemisphere temporal distributions. Mean scalp-topographies were centered over the left posterior temporal lobe (**Figure 1A**) with a similar topography over the right hemisphere (**Figure 2A**). For both clusters, log spectra collapsed across cluster ICs revealed distinct spectral peaks at ∼10 Hz (**Figures 1B** and **2B**) and ECD locations within the left and right posterior temporal lobes with an average dipole location at Taliarach coordinates [(*x,y,z*) −58,−36,8] in the left hemisphere and [(*x,y,z*) 61,−34,5] in the right hemisphere (**Figures 1C** and **2C**). The residual variance not explained by the single dipole model was 8.33% for the left hemisphere and 9.97% in the right hemisphere, indicating that a single dipole model accounted for ∼90% of the variance in the scalp distribution. To evaluate the statistical significance of cluster source estimates, statistical comparisons relative to zero (i.e., no activation) were computed for all sensorimotor and posterior temporal scalp topographies in the sLORETA statistical module (Grin-Yatsenko et al., 2010). A paired *t*-test was carried out for frequencies between 0.5 and 40 Hz (159 frames) with the smoothing parameter set to 1 (single common variance for all variables), using

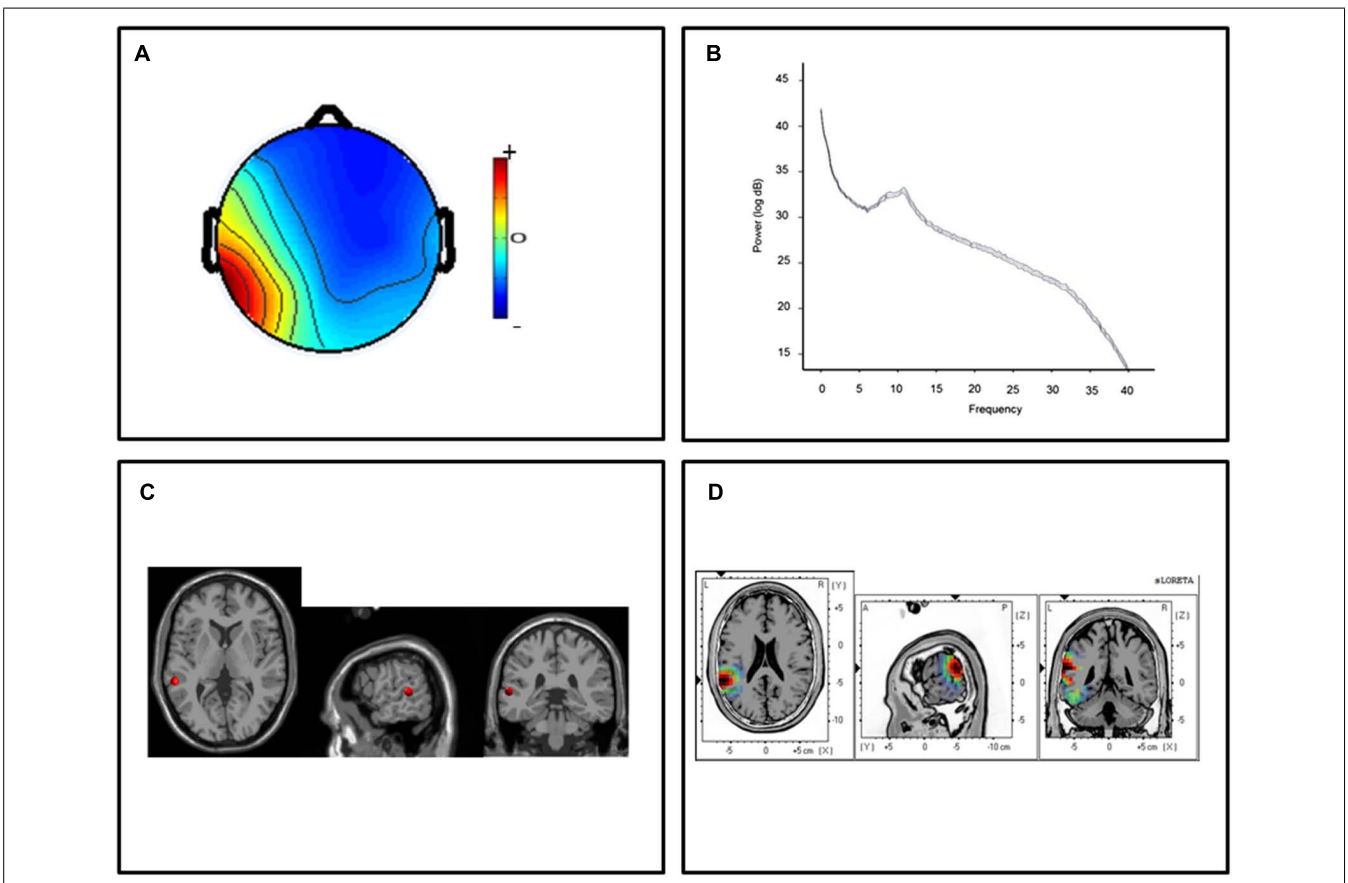


FIGURE 1 | Cluster results for the left-hemisphere α component. **(A)** Mean scalp potential distribution (W^{-1}) scaled to RMS microvolts and individual scalp distributions for each participant. **(B)** Mean spectra of the component across cluster ICs. **(C)** Average equivalent current dipole location, and **(D)** maximum current source density voxels (*t*-values) with greater values in darker colors and smaller values in lighter colors (NIH Micro template; at $p < 0.01$ corrected for multiple comparisons).

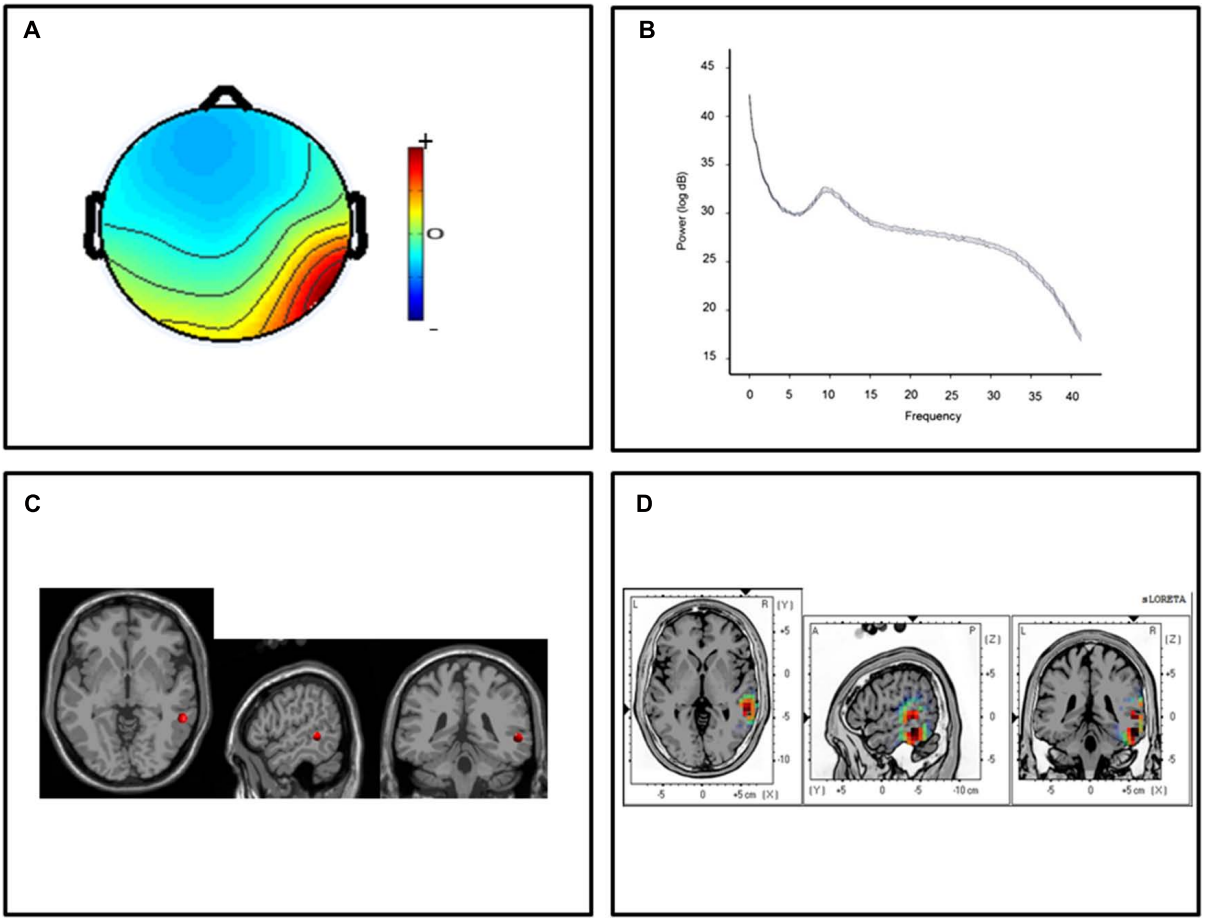


FIGURE 2 | Cluster results for the right-hemisphere α component. **(A)** Mean scalp potential distribution (W^{-1}) scaled to RMS microvolts and individual scalp distributions for each participant; **(B)** mean spectra of spectra of the component across cluster ICs. **(C)** Average equivalent current dipole location, and **(D)** maximum current source density voxels (*t*-values) with greater values in darker colors and smaller values in lighter colors (NIH Micro template; at $p < 0.01$ corrected for multiple comparisons).

5000 random permutations yielding corrected *t*-value threshold for all 6235 voxels in the sLORETA solution space. For temporal lobe clusters, a paired test revealed significant voxels at $p < 0.01$ in a region extending from the middle temporal gyrus to the parietal-temporal boundary with maximum CSD estimates at Taliarach [$t = 1.57(x,y,z)$ −64,−45,18] in the left hemisphere and Taliarach [$t = 2.07$ (x,y,z)−55,−41,16] in the right (**Figures 1D** and **2D**).

The characteristics of sensorimotor clusters are discussed in Bowers et al. (2013) and are consistent with well-known spectral and spatial features of the sensorimotor μ rhythm (Hari, 2006). The only difference between this analysis and that for the previous study is the head model used. The current study used a more realistic BEM whereas the previous study used a less realistic spherical model. Use of the BEM model resulted in a slightly more anterior mean dipole location in the left and right hemispheres at Taliarach coordinates [(x,y,z) −50,−11,33] for the left and [(x,y,z) 45,−16,43] for the right. The distributed solution (sLORETA) showed that the highest CSD estimates were distributed over the central sulcus in both the left Taliarach [$(x,y,z$ −45,−18,42)] and right hemispheres Taliarach [(x,y,z) 40,−16,61].

TEMPORAL LOBE CLUSTERS (α): ERSPs AND ITCs

Mean ERSP (**Figure 3**) ITC values (**Figure 5**) across subjects and conditions are shown in a time–frequency map with corrected significance values for condition in a separate map. Non-significant values are depicted in green and significant values are depicted in color from orange for weaker values to red for stronger values (pFDR < 0.10 to pFDR < 0.001). A repeated measures ANOVA design with the factor condition (1×7) revealed no significant differences for the number of trials submitted between conditions ($F = 0.92$, $p = 0.48$). The initial permutation analysis (1×7) revealed significant ERSPs in the 8–30 Hz range (alpha/beta) in the left α cluster and in the same range for the right hemisphere cluster corrected across the entire time–frequency matrix (pFDR < 0.05; 35×105; see **Figure 3**). Significant time–frequency values were found in the time-periods prior to, during, and after stimulus onset with peak event-related decreases in spectral power (i.e., ERD) in the time period after stimulus offset. The same statistical procedure (1×7; 32×105) was applied to the ITC dependent measure and showed significant ITCs commensurate with stimulus onset, with the strongest values extending to 800 ms following

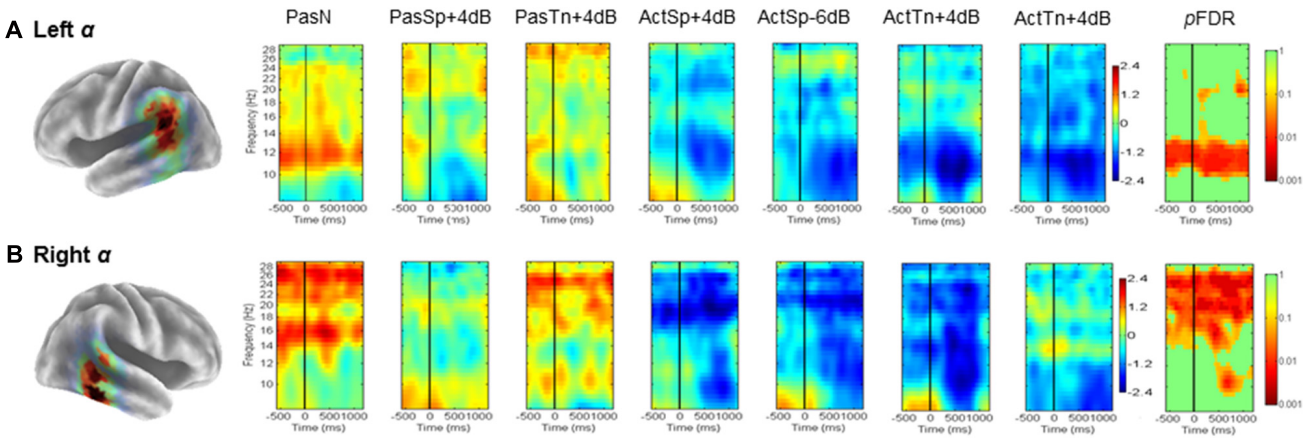


FIGURE 3 | Mean left and right hemisphere α time–frequency ERSPs (event-related spectral perturbations). ERSPs in root-mean-square decibel units as a function of condition (1 × 7) in the left **(A)** and right hemispheres **(B)**. FDR corrected *p*-values indicating significant effects in the beta (13–30Hz) and alpha ranges (8–13Hz) in the left **(A)** and right hemispheres **(B)**. Non-significant values are colored green, with significant values shown in orange and red. Event-related decreases in spectral power are indicated in blue (2.5) and increases are indicated in red (2.5).

stimulus onset in both left and right hemisphere component clusters.

To determine the sources of condition effects, first paired *t*-tests were used to compare each condition to the passive noise baseline (PasN). To test the initial hypotheses regarding ERSPs, the time periods before, during, and after stimulus onset were of interest and thus all subsequent analyses were restricted to the equal 600 ms time intervals prior to, during, and following stimulus onset (i.e., –600 to 1200 ms) prior to the cued response. First, for the ERSP dependent, paired comparisons to PasN revealed that only active conditions were associated with significant alpha suppression relative to PasN (pFDR < 0.05; 35 × 92). Planned comparisons designed to investigate task performance related effects showed no significant differences between correct trials in the ActSp + 4 dB condition and chance trials in the ActSp − 6 dB condition (pFDR > 0.05; 35 × 92) in either the left or right hemisphere cluster. The same comparison for ActTn + 4 dB and ActTn − 18 dB showed no significant difference in either hemisphere. As such, suppression in the alpha and beta frequencies was generally associated with active tasks demands but not with behavioral performance.

Second, for analysis of the ITC dependent all conditions were first compared with the PasN baseline in the left and right hemispheres in the time period from stimulus onset to 800 ms following stimulus onset in the 3–9 Hz range. Paired comparisons showed that passive conditions in both hemispheres were associated with phase reset relative to PasN (pFDR < 0.05; 28 × 41). A comparison of active conditions to baseline revealed at significant effect for the ActSp + 4 dB and ActTn + 4 dB conditions. A comparison of correct trials in the ActSp + 4 dB condition with chance trials in the ActSp − 6 dB condition showed a brief significant difference from 400 to 600 ms following stimulus onset in the left hemisphere. The same comparison in the right hemisphere showed no significant difference for speech trials or for correct trials in the ActTn + 4 dB condition compared to the ActTn − 18 dB condition.

Given that both active correct and passive conditions were associated with higher ITC magnitude relative to the passive noise baseline (PasN), it was unclear whether phase reset in the active conditions was due to active task performance or the quality of bottom-up acoustic information. In other words, as both stimulus types were presented at the same SNR, to determine whether active task demands or bottom-up acoustic information accounted for increases in ITC magnitude, the active and passive conditions were compared. No significant differences at pFDR > 0.05 were observed, suggesting that active task performance was not associated with increases in ITC magnitude in the left temporal lobe cluster.

SENSORIMOTOR CLUSTERS (μ): ERSPs AND ITCs

Mean ERSP values for correct and chance trials for the left hemisphere clusters are shown in **Figure 4**. Mean ITC values across conditions are shown in a time–frequency map with FDR corrected significance values for significant condition effects in a separate map (**Figure 5**). ERSPs in the sensorimotor clusters as reported in Bowers et al. (2013) were associated with significant suppression in the traditional beta range prior to, during and following stimulus onset. The only performance related effect was just following stimulus offset in the left hemisphere cluster for the active syllable discrimination task only (shaded region in **Figure 4**). For the analysis of sensorimotor ITCs, the initial permutation analysis adopting a repeated measures ANOVA design (1 × 7) revealed significant ITCs (pFDR < 0.05; 32 × 105) in the 3–9 Hz range. Paired comparisons to PasN revealed significant differences in both hemispheres in the ActSp + 4 dB, ActSp − 6 dB, and ActTn + 4 dB conditions only (pFDR < 0.05; 28 × 41). No significant differences in either hemisphere were associated with passive conditions. As such, unlike the temporal lobe clusters, the sensorimotor clusters were associated with increases in ITC magnitude related to active task performance only. However, it is worth noting that passive conditions were associated with individual variability relative to the silent recording interval ($p < 0.05$

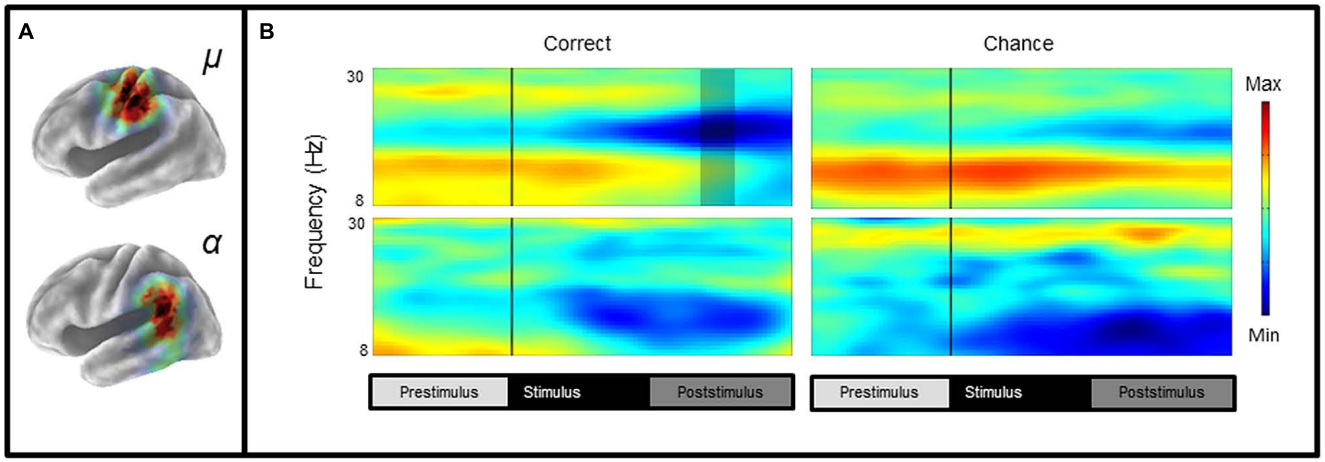


FIGURE 4 | Mean left hemisphere μ and α time-frequency ERSPs (event-related spectral perturbations) for correct and chance trials. **(A)** sLORETA images showing significant values in the sensorimotor and the temporal lobe regions, **(B)** ERSPs showing significant alpha and beta suppression in the active conditions for correct and chance trials. As reported in Bowers et al. (2013), the only performance related difference is in the time period following stimulus offset over the sensorimotor cortex (shaded region).

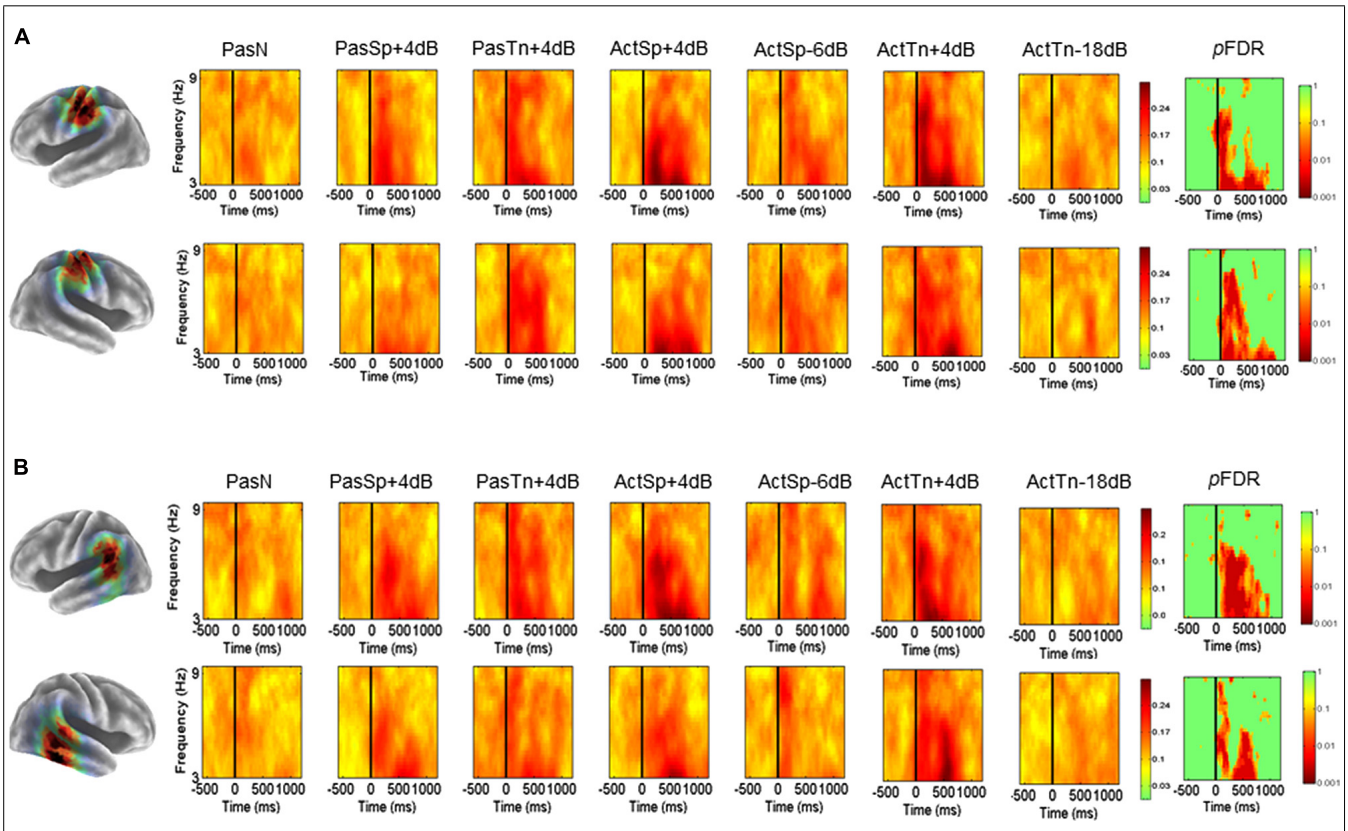


FIGURE 5 | Mean left and right hemisphere sensorimotor (μ) and temporal (α) ITCs. **(A)** Mean ITCs for the left and right sensorimotor component clusters as a function of condition with *p*-values corrected for false discovery rate in a separate map (non-significant values indicated in green). **(B)** Mean ITCs for the left and right posterior temporal component clusters as a function of condition with *p*-values corrected for false discovery rate in a separate map (non-significant values indicated in green).

uncorrected), suggesting that some participants presented with phase reset in passive conditions but the overall results did not survive the conservative correction for false discovery. For the left hemisphere cluster, performance related tests showed that correct trials in the ActSp + 4 dB conditions were significantly different from chance trials in the ActSp − 6 dB condition in the time period from 200 to 400 ms following stimulus onset ∼200 ms prior to the difference observed in the temporal lobe cluster (**Figures 6** and **7**). A comparison of correct ActTn + 4 dB trials with chance ActTn − 18 dB showed significant differences throughout stimulus presentation in both the left and right hemispheres.

DISCUSSION

The current analysis of event-related EEG in speech and non-speech discrimination investigated how hypothesized oscillatory mechanisms over the posterior temporal lobes function in time relative to those recorded over the sensorimotor cortex in a speech and non-speech discrimination task. The first aim was to demonstrate that alpha-like component clusters over the posterior temporal lobes are associated spectral suppression along shared at beta and alpha frequencies surrounding and during stimulus events. The second aim of the analysis was to investigate how oscillatory rhythms shared between sensorimotor and temporal lobe clusters vary depending on task and the quality of bottom-up acoustic input (i.e., correct relative to chance trials).

First, in accordance with our initial hypotheses, an IC cluster was found with a topography over the posterior temporal lobe characterized by mean peak spectra at ∼10 Hz and source estimates ranging from the posterior superior temporal sulcus to

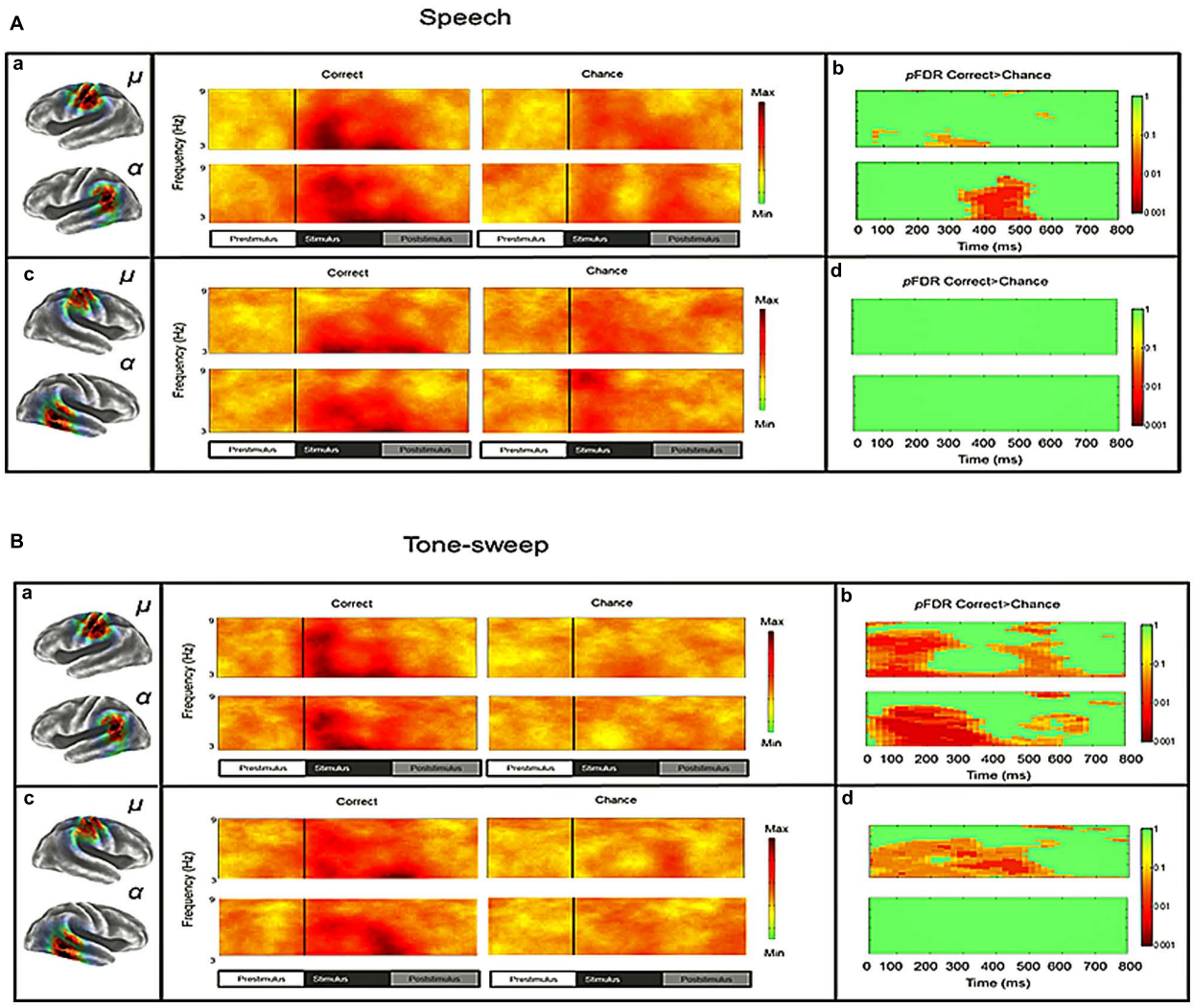


FIGURE 6 | Mean ITCS for correct and chance trials as a function of stimulus type and performance level for left and right sensorimotor and temporal clusters in the syllable discrimination condition. **(A)** Mean ITC values for left and right clusters in the syllable discrimination condition with significant differences between correct and chance trials depicted in a separate map (non-significant values in green). **(B)** Mean ITC values for left and right hemisphere clusters in the tone-sweep discrimination condition with significant differences between correct and chance trials depicted in a separate map (non-significant values in green).

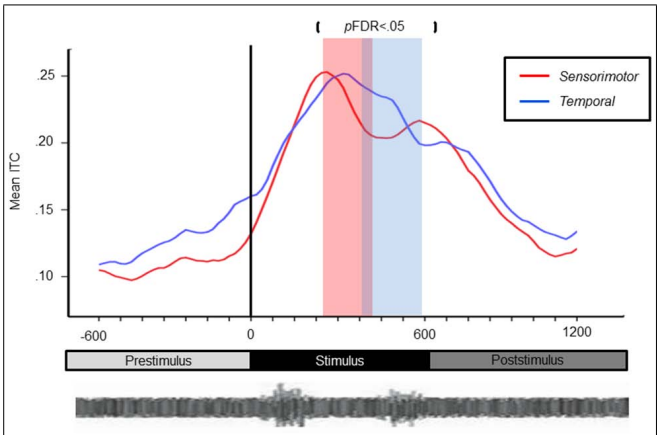


FIGURE 7 | Line chart depicting earlier peak responses in the left-sensorimotor clusters relative to temporal lobe clusters for correct syllable discrimination trials. Sensiromotor cluster shown in red with the temporal cluster shown in blue. Significance for the contrast correct > chance are marked in the red column for the sensorimotor cluster and blue for the temporal lobe cluster.

the parietotemporal boundary. sLORETA analysis showed that the greatest area of overlap between the posterior temporal scalp maps was in the posterior superior temporal gyrus (pSTG) near the parietal–temporal boundary. Second, alpha suppression was not different from baseline during passive tasks, yet active tasks were associated with alpha and beta suppression from the time period prior to stimulus onset to the time following stimulus offset. Third, activity in neither band was significantly related to correct relative to chance identification trials, suggesting a more general role in auditory attention not specifically related to perceptual performance. In accordance with initial hypotheses, the posterior temporal cluster was associated with phase reset predominantly in the delta–theta band reaching up into the low alpha band (here 3–9 Hz) that was also significantly related to perceptual performance for both control tone-sweep and syllable stimuli. Left lateralized performance related effects were found for syllables with a temporal integration window of ∼200 ms, while a left and right-hemisphere network was related to tone-sweep discrimination performance in the same time window. These findings are consistent with a class of oscillatory models that may be referred to collectively as "entrainment" theories predicting low frequency phase reset across the sensorimotor network thought be critical for parsing speech units (Schroeder and Lakatos, 2009; Giraud and Poeppel, 2012; Ghitza, 2013).

Critically, during active conditions power suppression in the beta range and phase reset in the in the delta–theta range occurred in sensorimotor components consistent with that in the posterior temporal lobe clusters, implicating entrained oscillatory mechanisms supporting task-related performance. Passive processing in the sensorimotor clusters was not found to be different from the PasN baseline, suggesting that robust phase reset in motor regions, unlike that in sensory regions, was not required during passive listening. However, during active processing, significant differences between active correct and chance trials were found earlier in the left sensorimotor cluster compared to those over the left posterior temporal lobe. As proposed by neurophysiological accounts of active processing (e.g., active sensing; analysis-by-synthesis; internal models), early efferent copies during active attention to syllable categorization may function to modulate processing focused on sensory events, resulting in increases in ITC magnitude consistent with the syllable unit in sensorimotor regions. Consistent with recent proposals (e.g., Arnal and Giraud, 2012), these findings suggest that shared mechanisms evident in locally synchronized rhythms contribute bidirectional information along oscillatory channels both from the top-down at higher frequencies and from the bottom-up at lower frequencies to mediate perceptual performance. In the discussion following, findings will first be framed within a synthesis of the literature regarding the accumulating evidence for an auditory cortical alpha rhythm and neuroimaging evidence for posterior temporal lobe activation in similar tasks. Second, dynamic time–frequency measures (i.e., ITCs and ERSPs) will be discussed relative to the functional role of sensorimotor integration in speech discrimination tasks. An overall interpretation of the findings will be discussed from a dynamic systems perspective.

POSTERIOR TEMPORAL ALPHA RHYTHMS

Despite the wide acceptance of well-established somatosensory and visual alpha rhythms, the presence of an independent auditory alpha rhythm has been met with skepticism (Weisz et al., 2011). However, an accumulating body of evidence implicates such a rhythm in auditory processing. The current finding of an IC cluster with an alpha-like signature over the posterior temporal lobes is consistent with an independent alpha rhythm in the auditory association areas and contributes to evidence supporting its role in speech processing. Two other studies using ICA of event-related EEG have detected an independent physiological process localized to the posterior temporal lobes related to auditory event-related potentials (Marco-Pallarés et al., 2005) and for a single subject showing alpha/beta suppression in audio-visual speech processing (Callan et al., 2001). Further, the existence of an independent auditory alpha rhythm with source estimates in the posterior temporal lobes is also broadly consistent with previous neuroimaging findings employing speech and non-speech discrimination tasks. Binder et al. (2004) found voxels in 13 of 18 subjects correlated with syllable identification accuracy were located in the left pSTG and right STS. The sources estimates reported here are also consistent with ECoG recordings over the pSTG (Chang et al., 2010) and are consistent with alpha suppression in the same region (Crone et al., 2001). A study using tone-sweeps characterized by a rapid transition similar to those used in the current study, also reported left lateralized effects in auditory association areas for both speech and non-speech signals containing a rapid temporal cue, further suggesting that auditory object identification generally relies on overlapping left and right hemisphere mechanisms for processing rapid acoustic transitions (Joanisse and Gati, 2003). Consistent with the asymmetric sampling in time hypothesis (AST), left and right hemisphere networks appear to share acoustic information at overlapping sampling rates, with a preference in left-hemisphere regions for integrating

rapidly transitioning cues (Poeppel, 2003). Future analyses using the current methodology might focus on how information at rapid (e.g., low gamma) and slow (delta–theta) rhythms are integrated in temporal lobe components.

SENSORIMOTOR INTEGRATION

The relative role of motor and auditory subsystems in resolving the inherent variability of the speech signal is controversial (Gallese et al., 2011). The current findings contribute to this debate by providing high-temporal resolution measures prior to, during, and following sensory events along oscillatory channels proposed to play an important functional role in perception and sensorimotor integration (Callan et al., 2010; Arnal and Giraud, 2012; Giraud and Poeppel, 2012; Obleser et al., 2012). First, the current findings support the conclusion that motor and higher order sensory subsystems function in different rhythmic modes for active relative to passive tasks. For measures of power, passive tasks were not significantly different from the PasN baseline in either sensorimotor or temporal clusters, while active tasks were associated with suppression at alpha and beta frequencies. Bottom-up phase responsive mechanisms in higher-order auditory regions were driven by stimulus input, were phase responsive to acoustic stimulation generally, and were reduced to baseline levels when acoustic cues were severely degraded. Phase reset in sensorimotor regions was not robustly active during passive listening, was responsive to the task regardless of acoustic degradation level, and was differentially enhanced for speech relative to non-speech stimuli when sensory input supported task goals (i.e., correct trials). These findings suggest that active top-down mechanisms reflecting release from inhibition were recruited primarily due to active attention to task demands and were selective to expected input, whereas bottom-up sensory mechanisms were active during acoustic stimulation regardless of task or the auditory stimulus employed.

One caveat to the preceding conclusions is that it remains possible that greater degradation or ambiguity of acoustic cues might activate automatic phase reset in sensorimotor regions in a passive task (Osnes et al., 2011). A recent TMS study demonstrated automatic motor influences on auditory processing in during the presentation of acoustic cues in which speech stimuli were manipulated along an F1–F2 continuum (Möttönen et al., 2013). In both the Osnes et al. (2011) and Möttönen et al. (2013) studies, the ambiguity of acoustic cues for stimulus perception appears to have induced automatic activity in motor regions, a process that might be characterized as a feedforward mechanism connecting early sensory hypotheses with articulatory representations in motor regions to aid in resolving perceptual ambiguity. This mechanism may be contrasted with the role of the motor system when participants anticipate expected features of the auditory stimulus in the service of task goals, which are thought to be propagated backward in the cortical hierarchy via articulatory models (Callan et al., 2010; Hickok et al., 2011; Arnal and Giraud, 2012). Thus, taken in the context of evidence from a range of active and passive tasks, the current results implicate more than one function for sensory and motor subsystems in speech discrimination varying with attention to auditory stimuli and the quality of acoustic information conveyed by the stimulus.

Second, the current analysis suggests that, during tasks requiring active discrimination, left hemisphere sensorimotor systems have an earlier performance related effect on delta–theta phase reset relative to left-hemisphere temporal lobe clusters. Although non-speech rapid-auditory processing activated the same left-hemisphere sensorimotor network, no performance related time differential between sensorimotor and temporal activity was observed. This finding is consistent with explanations proposed by a number of research groups to account for how sensorimotor experience (i.e., procedural knowledge) with speech production could have a modulatory influence on speech discrimination. According to these proposals, early articulatory models prior to acoustic onset provide general predictions about the most likely upcoming spectro-temporal features. Early beta suppression prior to sensory input can be explained as an early internal model related to active attention to task demands. During sensory input, bottom-up information induced by stimulus onset and shared along delta–theta channels is modulated so that earlier activity consistent with a speech specific internal model occurs in sensorimotor regions with later activity in temporal lobe regions critical for categorization. This explanation is consistent with another recent study employing a duplex perception paradigm. In that study, ICA of both hemodynamic and EEG signals demonstrated early activity in the left lateralized somatomotor regions 250 ms prior to those the in the pSTG during active phonological processing (Liebenthal et al., 2013). However, in the current study, given that performance related effects were driven by the quality of bottom-up input and differences in phase reset occurred only during active tasks in sensorimotor clusters, somatomotor processing represents an adaptation to task requirements in the active condition. Consistent with lesion evidence suggesting that the motor system plays a secondary role in speech processing, these findings support a model weighted toward bottom-up sensory analysis with top-down modulatory influence from sensorimotor regions (Hickok et al., 2011; Schwartz et al., 2012; Bowers et al., 2013; Möttönen et al., 2013) similar to recent revivals of the theory of analysis-by-synthesis (Poeppel and Monahan, 2011).

DYNAMIC OSCILLATORY MODELS

A wide range of explanations have been proposed to account for the activation of sensorimotor networks in the context of active and passive listening, including a role in attention/working memory (LoCasto et al., 2004; Szenkovits et al., 2012), covert rehearsal (Hickok and Poeppel, 2007; Hickok et al., 2011), a role in stimulus expectancy (Osnes et al., 2012), the resolution of ambiguous acoustic cues (Callan et al., 2010; Osnes et al., 2011), and articulatory selective attention implemented via efferent internal models associated with speech production (Callan et al., 2010; Hickok et al., 2011). An explanation compatible with multiple roles for sensorimotor networks in different contexts can be derived from dynamic theories of cognition (Engel et al., 2001; Engel and Fries, 2010). According to dynamic oscillatory theories, degenerate mappings between local neuronal populations may function flexibly in different global oscillatory patterns to achieve the same perceptual outcomes (Engel et al., 2001). In general, dynamic theories predict that internally generated states of anticipation or expectancy result in large-scale coherence across regions known

as dynamic resonance. At the same time, local cell populations with specified receptive fields compete for stable resonant states reflecting a best match between bottom-up sensory input and internally generated predictions about upcoming sensory feature constellations.

A dynamic explanation suggests two distinct processing strategies may emerge for passive and active listening tasks. During passive listening, oscillatory dynamics appear to exist in a self-organized, coordinative state reflected primarily in low frequency oscillations in the sensorimotor network thought to be critical for categorical perception (Schroeder and Lakatos, 2009; Giraud and Poeppel, 2012). As would be expected for a passive task in which acoustic cues are clear, significant delta–theta enhancement was apparent only in higher order auditory cortices known to be involved in the categorization of bottom-up input. Given the behavioral results in the current study, it is unlikely that participants had difficulty discriminating between syllables in the passive condition, yet during the active condition a different processing strategy consistent with internal models emerged. In other words, an internally generated state related to attention to task demands induced a different pattern of oscillatory activity for what is most likely the same perceptual outcome. During goal-directed discrimination, oscillatory channels linked to auditory association areas via previous sensorimotor experience are simultaneously disinhibited prior to sensory input, with peak activity occurring in the sensorimotor cortex when bottom-up cues are sufficient to specify speech units for discrimination. A decrease in the left sensorimotor cortex occurs when bottom-up cues are not sufficient to support task goals, reflecting a mismatch between somatomotor predictions and spectro-temporal processing. As such, oscillatory activity in the left sensorimotor cortex may be characterized as speech selective component of goal-directed selective attention within the auditory dorsal auditory stream (Skipper et al., 2006; Callan et al., 2010; Hickok et al., 2011).

Although the current study suggests a role for sensorimotor representations during the performance of a syllable discrimination task, it remains unclear how sensorimotor predictions might function in real-world contexts. One possibility is that goal-directed attention to various features of the communicative signal might stabilize patterns of neural activity that would otherwise be unstable via shared mechanisms in global networks (Kelso, 2012). This notion might tentatively suggest that top-down influences in sensorimotor networks aid in generating stable percepts by modulating oscillatory phase dynamics at time-constants consistent with the syllable unit (Ghitza, 2013) with greater weight on auditory association or motor regions depending upon context (Skipper et al., 2006). This conjecture is defensible as recent evidence supports the conclusion that segmental properties of speech predict word recognition, suggesting that each segment is involved in computing the next segment (Gagnepain et al., 2012). However, it is an open question whether or not motor systems involved in a speech discrimination task also play a functional role in, for example, a conversation in a crowded room. It is likely that the same mechanism would be in selective competition with other entrained top-down mechanisms (e.g., the ventral stream) involved in linguistic and gestural analysis (Skipper et al., 2006, 2009). In more naturalistic contexts, speech units occur at predictable temporal intervals and are accompanied by a host of linguistic features and gestures known to influence perception and comprehension (Skipper et al., 2009; Morillon et al., 2010; Arnal and Giraud, 2012). As such, a better understanding of how the motor system functions in speech processing in relation to the ventral and dorsal streams might be achieved by manipulating predictive mechanisms in more naturalistic contexts.

LIMITATIONS AND CONCLUSIONS

An important limitation of the spatial estimates in the current study is the inability to determine the responses of subregions within the temporal lobe or sensorimotor distribution due to the inherent low resolution of sLORETA estimates. The left-hemisphere region implicated in the current study suggests greater CSD estimates in a heteromodal region known to be involved in mediating sensorimotor transformations during speech production (Hickok and Poeppel, 2007), suggesting that the current distribution may have been pulled toward this region due to the activation of sensorimotor integration processes. Although preliminary evidence would suggest that conventional random effects analysis of hemodynamic measures is associated with dipole models of IC scalp topographies, few studies have investigated such a relationship (Debener et al., 2005). Given the reported inverse relationship between BOLD measures and alpha/beta suppression (Yuan et al., 2010), conceivably the signal processing approach used in this study may be used with simultaneous high-density EEG, individual participant MR head models, and more spatially precise hemodynamic methods to investigate subregions within the sensorimotor networks and how they are related to alpha and beta suppression.

A second limitation is that while sensorimotor and temporal lobe clusters were associated with activity along shared oscillatory channels and condition differences implicate competition in the sensorimotor network, high spatial, and time–frequency resolution dynamic causal models (DCM) would be required to explicitly test how sensorimotor networks directionally vary their connection weights (Chen et al., 2008). Potentially, connectivity models using regions of interest indicated by source models of ICA topographies could be used to test the hypothesis that cortical patches vary their connection weights in time along relevant frequency channels (Chen et al., 2012). A third limitation is that, for the sake of simplicity, the role of gamma rhythms was not explored here. A specific role has been proposed for gamma oscillations in propagating feedforward error when bottom-up features are at odds with predictive internal models, suggesting that they may play an important functional role via interaction with alpha, beta, and delta–theta oscillations (Arnal and Giraud, 2012). Future studies should also explore the role of low gamma rhythms in perceptual tasks along with lower frequency components of the signal.

To our knowledge, the current study is the first to implicate simultaneously measured phase reset and power suppression of sensory and sensorimotor rhythms in a discrimination task commonly employed in neuroimaging experiments. The study suggests that sensorimotor and auditory rhythms are shared when participants are engaged in goal-directed listening and are distinct from those involved in passive listening. The study provides further

evidence for a speech selective role of the left sensorimotor cortex along beta and delta–theta channels consistent with a role in articulatory selective attention (Callan et al., 2010; Hickok et al., 2011). The study provides initial support for the predictions of recent oscillatory frameworks in which beta and delta–theta channels are proposed to play a role in perception depending on context (Arnal and Giraud, 2012). Further, consistent with dynamic oscillatory accounts, this study suggests that while auditory and sensorimotor regions share processing along the same oscillatory channels, selective enhancement in the two respective regions is dependent on the task and quality of sensory input, implicating competition between locally synchronized regions along the same oscillatory channels. We suggest that the importance of using EEG to provide evidence for these mechanisms is that the recording method has potential for use in speech and hearing clinics where other neuroimaging methods are often unavailable. As a number of communication disorders are also associated with spectro-temporal processing deficits, an understanding of how dynamic oscillatory systems compensate for changing informational demands on a millisecond timescale may be critical to an understanding of how perceptual processes succeed or fail in individuals with speech, hearing, and language deficits.

REFERENCES

Adank, P. (2012). Design choices in imaging speech comprehension: an activation likelihood estimation (ALE) meta-analysis. *Neuroimage* 63, 1601–1613. doi: 10.1016/j.neuroimage.2012.07.027

Alho, J., Sato, M., Sams, M., Schwartz, J. L., Tiitinen, H., and Jääskeläinen, I. P. (2012). Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. *Neuroimage* 60, 1937–1946. doi: 10.1016/j.neuroimage.2012.02.011

Arnal, L. H., and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.* 57, 289–300. doi: 10.2307/2346101

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., and Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nat. Neurosci.* 7, 295–301. doi: 10.1038/nn1198

Bland, B. H., and Oddie, S. D. (2001). Theta band oscillation and synchrony in the hippocampal formation and associated structures: the case for its role in sensorimotor integration. *Behav. Brain Res.* 127, 119–136. doi: 10.1016/S0166-4328(01)00358-8

Bowers, A., Saltuklaroglu, T., Harkrider, A., and Cuellar, M. (2013). Suppression of the μ rhythm during speech and non-speech discrimination revealed by independent component analysis: implications for sensorimotor integration in speech processing. *PLoS ONE* 8:e72024. doi: 10.1371/journal.pone.0072024

Callan, D., Callan, A., Gamez, M., Sato, M., and Kawato, M. (2010). Premotor cortex mediates perceptual performance. *Neuroimage* 51, 844–858. doi: 10.1016/j.neuroimage.2010.02.027

Callan, D. E., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: a single-sweep EEG case study. *Brain Res. Cogn. Brain Res.* 10, 349–353. doi: 10.1016/S0926-6410(00)00054-9

Callan, D. E., Kent, R. D., Guenther, F. H., and Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *J. Speech Lang. Hear. Res.* 43, 721–736. doi: 10.1044/jslhr.4303.721

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Chen, C. C., Kiebel, S. J., and Friston, K. J. (2008). Dynamic causal modelling of induced responses. *Neuroimage* 41, 1293–1312. doi: 10.1016/j.neuroimage.2008.03.026

Chen, J.-L., Ros, T., and Gruzelier, J. H. (2012). Dynamic changes of ICA-derived EEG functional connectivity in the resting state. *Hum. Brain Mapp.* 34, 852–868. doi: 10.1002/hbm.21475

Crone, N. E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* 112, 565–582. doi: 10.1016/S1388-2457(00)00545-9

Cuellar, M., Bowers, A., Harkrider, A. W., Wilson, M., and Saltuklaroglu, T. (2012). Mu suppression as an index of sensorimotor contributions to speech processing: evidence from continuous EEG signals. *Int. J. Psychophysiol.* 85, 242–248. doi: 10.1016/j.ijpsycho.2012.04.003

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Von Cramon, D. Y., and Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J. Neurosci.* 25, 11730–11737. doi: 10.1523/JNEUROSCI.3286-05.2005

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., and Makeig, S. (2012). Independent EEG sources are dipolar. *PLoS ONE* 7:e30135. doi: 10.1371/journal.pone.0030135

Doelling, K., Arnal, L., Ghitza, O., and Poeppel, D. (2013). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85, 761–768. doi: 10.1016/j.neuroimage.2013.06.035

Driver, J., and Frith, C. (2000). Shifting baselines in attention research. *Nat. Rev. Neurosci.* 1, 147–148. doi: 10.1038/35039083

Engel, A. K., Fries, P., and Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. *Nat. Rev. Neurosci.* 2, 704–716. doi: 10.1038/35094565

Engel, A. K., and Fries, P. (2010). Beta-band oscillations – signalling the status quo? *Curr. Opin. Neurobiol.* 20, 156–165. doi:10.1016/j.conb.2010.02.015

Ernst, S. M. A., Verhey, J. L., and Uppenkamp, S. (2008). Spatial dissociation of changes of level and signal-to-noise ratio in auditory cortex for tones in noise. *Neuroimage* 43, 321–328. doi: 10.1016/j.neuroimage.2008.07.046

Fellinger, R., Klimesch, W., Gruber, W., Freunberger, R., and Doppelmayr, M. (2011). Pre-stimulus alpha phase-alignment predicts P1-amplitude. *Brain Res. Bull.* 85, 417–423. doi: 10.1016/j.brainresbull.2011.03.025

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci.* 9, 474–480. doi: 10.1016/j.tics.2005.08.011

Frith, C. D., Blakemore, S.-J., and Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. *Brain Res. Rev.* 31, 357–363. doi: 10.1016/S0165-0173(99)00052-1

Gagnepain, P., Henson, R. N., and Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Curr. Biol.* 22, 615–621. doi: 10.1016/j.cub.2012.02.015

Gallese, V., Gernsbacher, M. A., Heyes, C., Hickok, G., and Iacoboni, M. (2011). Mirror neuron forum. *Perspect. Psychol. Sci.* 6, 369–407. doi: 10.1177/1745691611413392

Ghitza, O. (2013). The theta-syllable: a unit of speech information defined by cortical function. *Front. Psychol.* 4:138. doi: 10.3389/fpsyg.2013.00138

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., and Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* 56, 1127–1134. doi: 10.1016/j.neuron.2007.09.038

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063

Grabski, K., Tremblay, P., Gracco, V. L., Girin, L., and Sato, M. (2013). A mediating role of the auditory dorsal pathway in selective adaptation to speech: a state-dependent transcranial magnetic stimulation study. *Brain Res.* 1515, 55–65. doi: 10.1016/j.brainres.2013.03.024

Grin-Yatsenko, V. A., Baas, I., Ponomarev, V. A., and Kropotov, J. D. (2010). Independent component approach to the analysis of EEG recordings at

early stages of depressive disorders. *Clin. Neurophysiol.* 121, 281–289. doi: 10.1016/j.clinph.2009.11.015

Hari, R. (2006). Action–perception connection and the cortical mu rhythm. *Prog. Brain Res.* 159, 253–260. doi: 10.1016/S0079-6123(06)59017-X

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Jasper, H. A. (1958). The ten–twenty system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* 10, 371–375.

Joanisse, M. F., and Gati, J. S. (2003). Overlapping neural regions for processing rapid temporal cues in speech and nonspeech signals. *Neuroimage* 19, 64–79. doi: 10.1016/S1053-8119(03)00046-6

Kelso, J. S. (2012). Multistability and metastability: understanding dynamic coordination in the brain. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 906–918. doi: 10.1098/rstb.2011.0351

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., and Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neurosci. Lett.* 244, 73–76. doi: 10.1016/S0304-3940(98)00122-0

Lee, T. W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput.* 11, 417–441. doi: 10.1162/089976699300016719

Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., and Desai, A. (2013). Neural dynamics of phonological processing in the dorsal auditory stream. *J. Neurosci.* 33, 15414–15424. doi:10.1523/JNEUROSCI.1511-13.2013

LoCasto, P. C., Krebs-Noble, D., Gullapalli, R. P., and Burton, M. W. (2004). An fMRI investigation of speech and tone segmentation. *J. Cogn. Neurosci.* 16, 1612–1624. doi: 10.1162/0898929042568433

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010. doi:10.1016/j.neuron.2007.06.004

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behav. Brain Sci.* 21, 499–511. doi: 10.1017/S0140525X98001265

Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004). Mining event-related brain dynamics. *Trends Cogn. Sci.* 8, 204–210. doi: 10.1016/j.tics.2004.03.008

Marco-Pallarés, J., Grau, C., and Ruffini, G. (2005). Combined ICA-LORETA analysis of mismatch negativity. *Neuroimage* 25, 471–477. doi: 10.1016/j.neuroimage.2004.11.028

Massaro, D. W. (1974). Perceptual units in speech recognition. *J. Exp. Psychol.* 102, 199–208. doi: 10.1037/h0035854

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2010). ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* doi: 10.1111/j.1469-8986.2010.01061.x [Epub ahead of print].

Morillon, B., Lehongre, K., Frackowiak, R. S., Ducorps, A., Kleinschmidt, A., Poeppel, D., et al. (2010). Neurophysiological origin of human brain asymmetry for speech and language. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18688–18693. doi: 10.1073/pnas.1007189107

Möttönen, R., Dutton, R., and Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cereb. Cortex* 23, 1190–1197. doi: 10.1093/cercor/bhs110

Möttönen, R., and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825. doi: 10.1523/jneurosci.6018-08.2009

Möttönen, R., and Watkins, K. E. (2012). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology* 26, 1103–1118. doi: 10.1080/02687038.2011.619515

Obleser, J., Herrmann, B., and Henry, M. J. (2012). Neural oscillations in speech: don't be enslaved by the envelope. *Front. Hum. Neurosci.* 6:250. doi: 10.3389/fnhum.2012.00250

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Osnes, B., Hugdahl, K., Hjelmervik, H., and Specht, K. (2012). Stimulus expectancy modulates inferior frontal gyrus and premotor cortex activity in auditory perception. *Brain Lang.* 121, 65–69. doi: 10.1016/j.bandl.2012.02.002

Osnes, B., Hugdahl, K., and Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *Neuroimage* 54, 2437–2445. doi: 10.1016/j.neuroimage.2010.09.078

Onton, J., and Makeig, S. (2006). "Information-based modeling of event-related brain dynamics," in *Event-Related Dynamics of Brain Oscillations,* eds C. Neuper and W. Klimesch (Elsevier), 99–120. doi: 10.1016/S0079-6123(06)59007-7

Oostenveld, R., and Oostendorp, T. F. (2002). Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. *Hum. Brain Mapp.* 17, 179–192. doi: 10.1002/hbm.10061

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find. Exp. Clin. Pharmacol.* 24(Suppl. D), 5–12.

Peelle, J. E., and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* 3:320. doi: 10.3389/fpsyg.2012.00320

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time." *Speech Commun.* 41, 245–255. doi: 10.1016/S0167-6393(02)00107-3

Poeppel, D., and Monahan, P. J. (2011). Feedforward and feedback in speech perception: revisiting analysis by synthesis. *Lang. Cogn. Process.* 26, 935–951. doi: 10.1080/01690965.2010.493301

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Rauschecker, J. P., and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Sato, M., Tremblay, P., and Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7. doi: 10.1016/j.bandl.2009.03.002

Schroeder, C. E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* 32, 9–18. doi: 10.1016/j.tins.2008.09.012

Schwartz, J.-L., Basirat, A., Ménard, L., and Sato, M. (2012). The perception-for-action-control theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguistics* 25, 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Siegel, M., Donner, T. H., and Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nat. Rev. Neurosci.* 13, 121–134. doi: 10.1038/nrn3137

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Curr. Biol.* 19, 661–667. doi: 10.1016/j.cub.2009.02.051

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2006). "Lending a helping hand to hearing: another motor theory of speech perception," in *Action to Language Via the Mirror Neuron System,* ed. M. A. Arbib (Cambridge, MA: Cambridge University Press), 250–286.

Specht, K. (2014). Neuronal basis of speech comprehension. *Hear. Res.* 307, 121–135. doi: 10.1016/j.heares.2013.09.011

Stevens, K. N., and Halle, M. (1967). "Remarks on analysis by synthesis and distinctive features," in *Models for the Perception of Speech and Visual Form* (Cambridge, MA: MIT Press), 88–102.

Szenkovits, G., Peelle, J. E., Norris, D., and Davis, M. H. (2012). Individual differences in premotor and motor recruitment during speech perception. *Neuropsychologia* 50, 1380–1392. doi: 10.1016/j.neuropsychologia.2012.02.023

Towle, V. L., Bolaños, J., Suarez, D., Tan, K., Grzeszczuk, R., Levin, D. N., et al. (1993). The spatial location of EEG electrodes: locating the best-fitting sphere relative to cortical anatomy. *Electroencephalogr. Clin. Neurophysiol.* 86, 1–6. doi: 10.1016/0013-4694(93)90061-Y

Venezia, J. H., Saberi, K., Chubb, C., and Hickok, G. (2012). Response bias modulates the speech motor system during syllable discrimination. *Front. Psychol.* 3:157. doi: 10.3389/fpsyg.2012.00157

Weisz, N., Hartmann, T., Müller, N., Lorenz, I., and Obleser, J. (2011). Alpha rhythms in audition: cognitive and clinical perspectives. *Front. Psychol.* 2:73. doi: 10.3389/fpsyg.2011.00073

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi:10.1038/nn1263

Yuan, H., Liu, T., Szarkowski, R., Rios, C., Ashe, J., and He, B. (2010). Negative covariation between task-related responses in alpha/beta-band activity and BOLD in human sensorimotor cortex: an EEG and fMRI study of motor imagery and movements. *Neuroimage* 49, 2596–2606. doi: 10.1016/j.neuroimage.2009.10.028

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 24 January 2014; accepted: 07 April 2014; published online: 07 May 2014.
Citation: Bowers AL, Saltuklaroglu T, Harkrider A, Wilson M and Toner MA (2014) Dynamic modulation of shared sensory and motor cortical rhythms mediates speech and non-speech discrimination performance. Front. Psychol. ***5****:366. doi: 10.3389/fpsyg.2014.00366*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 May 2014
doi: 10.3389/fpsyg.2014.00389

Multisensory and modality specific processing of visual speech in different regions of the premotor cortex

***Daniel E. Callan*[1,2]*, *Jeffery A. Jones*[3] *and Akiko Callan*[1,2]**

[1] *Center for Information and Neural Networks, National Institute of Information and Communications Technology, Osaka University, Osaka, Japan*
[2] *Multisensory Cognition and Computation Laboratory Universal Communication Research Institute, National Institute of Information and Communications Technology, Kyoto, Japan*
[3] *Psychology Department, Laurier Centre for Cognitive Neuroscience, Wilfrid Laurier University, Waterloo, ON, Canada*

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Anna J. Simmonds, Imperial College London, UK
Howard Charles Nusbaum, The University of Chicago, USA

****Correspondence:***
Daniel E. Callan, Center for Information and Neural Networks, National Institute of Information and Communications Technology, Osaka University, 1-4 Yamadaoka, Osaka 565-0871, Japan
e-mail: dcallan@nict.go.jp

Behavioral and neuroimaging studies have demonstrated that brain regions involved with speech production also support speech perception, especially under degraded conditions. The premotor cortex (PMC) has been shown to be active during both observation and execution of action ("Mirror System" properties), and may facilitate speech perception by mapping unimodal and multimodal sensory features onto articulatory speech gestures. For this functional magnetic resonance imaging (fMRI) study, participants identified vowels produced by a speaker in audio-visual (saw the speaker's articulating face and heard her voice), visual only (only saw the speaker's articulating face), and audio only (only heard the speaker's voice) conditions with varying audio signal-to-noise ratios in order to determine the regions of the PMC involved with multisensory and modality specific processing of visual speech gestures. The task was designed so that identification could be made with a high level of accuracy from visual only stimuli to control for task difficulty and differences in intelligibility. The results of the functional magnetic resonance imaging (fMRI) analysis for visual only and audio-visual conditions showed overlapping activity in inferior frontal gyrus and PMC. The left ventral inferior premotor cortex (PMvi) showed properties of multimodal (audio-visual) enhancement with a degraded auditory signal. The left inferior parietal lobule and right cerebellum also showed these properties. The left ventral superior and dorsal premotor cortex (PMvs/PMd) did not show this multisensory enhancement effect, but there was greater activity for the visual only over audio-visual conditions in these areas. The results suggest that the inferior regions of the ventral premotor cortex are involved with integrating multisensory information, whereas, more superior and dorsal regions of the PMC are involved with mapping unimodal (in this case visual) sensory features of the speech signal with articulatory speech gestures.

Keywords: audio-visual, premotor, multisensory, mirror system, fMRI, internal model

INTRODUCTION

Visual observation of gestural information available from a speaker's face improves speech perception, especially under noisy conditions (Sumby and Pollack, 1954; Grant and Braida, 1991; Callan et al., 2001, 2003). Speech gesture information, which consists of the biological motion of the various articulators (jaw, lips, tongue, larynx) that specify vocal tract shape, facilitates speech perception because of the direct relationship between vocal tract shape, speech acoustics, and the dynamic deformation of the skin of the face. Brain imaging studies suggest that the brain regions involved in the integration of multisensory information process gestural speech information to facilitate speech perception (Callan et al., 2003, 2004a,b; Skipper et al., 2007a,b). One means by which speech intelligibility may be enhanced by the addition of visual information is via brain regions that are involved in the multisensory integration process. Integration of temporally concordant information from multiple sensory channels (e.g., auditory and visual modalities) within specific brain regions, such as the superior temporal gyrus/sulcus (STG/S) in the case of audio-visual speech (Calvert et al., 2000; Callan et al., 2001, 2003; Sekiyama et al., 2003), results in enhanced neural activity that is greater than the combined activity in response to unimodal speech stimuli presented alone.

Another property of multisensory integration is the principle of inverse effectiveness, which asserts that multisensory enhancement is greatest under conditions in which unimodal stimuli elicit weak neural responses (e.g., due to subthreshold stimulation, noisy conditions; Wallace et al., 1992; Stein and Meredith, 1993). This multisensory enhancement effectively increases perceptual acuity and is maximized by temporally and spatially concordant stimulation of different sensory modalities (e.g., auditory and visual) (Stein and Meredith, 1993). The STG/S as well as the inferior frontal gyrus IFG/Broca's area have been shown to be involved in multisensory enhancement during perception of audio-visual speech in noise (Callan et al., 2001, 2003, 2004b; Alho et al., 2012).

Many researchers have proposed that speech intelligibility is enhanced by visual speech cues because the information available in the visible gestures activates motor representations that can be used to constrain auditory speech perception. Specifically, researchers hypothesize that certain brain regions internally model and simulate speech production and that these internal models are used to recover vocal tract shape information inherent in the speech signal (Callan et al., 2003, 2004a; Wilson and Iacoboni, 2006; Iacoboni and Wilson, 2006; Skipper et al., 2007a,b; Iacoboni, 2008; Poeppel et al., 2008; Rauschecker and Scott, 2009; Rauschecker, 2011). Internal models are a well-known concept in the motor control literature, and are believed to be used by the brain to simulate the input/output characteristics, or their inverses, of the motor control system (Kawato, 1999). In the case of speech, the forward and inverse mappings of the relationship between aspects of speech articulation and the acoustic features of speech output (as well as the orosensory and visual properties of speech) may be used to facilitate speech perception. Forward internal models predict the sensory (auditory, orosensory) consequences of the actions of speech articulation, whereas, inverse internal models determine the motor commands needed to articulate a desired sensory (auditory, orosensory) target. Callan et al. (2004a, 2010) suggested that the auditory consequences of internally simulated articulatory control signals (articulatory-auditory internal models for various phonemes) are used to constrain and facilitate speech perception under ambiguous conditions (e.g., speech perception in noisy environments, or the perception of non-native speech) through the competitive selection of the internal model that best matches the ongoing auditory signal. These internal models are thought to be instantiated in a network of speech motor regions that include the PMC and Broca's area, auditory processing regions STG/S, the IPL, and the cerebellum. Other researchers such as Rauschecker and Scott (2009) have discussed the use of forward and inverse auditory—articulatory mappings (utilizing principles of internal models) for speech perception and production, and have suggested that the IPL serves as an interface for matching of these mappings.

Several theories have proposed that speech perception uses aspects of speech production to extract phonetic information from sensory stimulation: Motor theory (Liberman et al., 1967), revised motor theory (Liberman and Mattingly, 1985; Liberman and Whalen, 2000), and various constructivist based theories (Callan et al., 2004a, 2010; Skipper et al., 2007a; Rauschecker and Scott, 2009; Rauschecker, 2011) including the Perception for Action Control Theory (PACT) (Schwartz et al., 2012). The observation of Mirror Neuron system like properties (active both during observation and execution of action) in Broca's area, the ventral inferior premotor cortex (PMvi) and the ventral superior and dorsal premotor cortex (PMvs/PMd), during speech production and perception has provided support for theories that propose a role for the motor system in speech perception (Callan et al., 2000a,b, 2006a,b, 2010; Wilson et al., 2004; Nishitani et al., 2005; Meister et al., 2007).

A number of studies have shown that these brain regions that appear to have Mirror Neuron system like properties, such as Broca's area and premotor cortex (PMC), respond to audio, visual, and audio-visual speech information (Campbell et al., 2001; Bernstein et al., 2002; Nishitani and Hari, 2002; Olson et al., 2002; Callan et al., 2003, 2004a,b; Paulesu et al., 2003; Calvert and Campbell, 2003; Ojanen et al., 2005; Skipper et al., 2005, 2007b; Alho et al., 2012; Dubois et al., 2012; Mashal et al., 2012). As well, the cerebellum has been shown to be involved in both perception and production of speech and is thought to instantiate processes related to internal models (Kawato, 1999; Imamizu et al., 2000; Callan et al., 2004a, 2007; Rauschecker, 2011; Tourville and Guenther, 2011; Callan and Manto, 2013). The objective of this study is to determine if these various brain regions (Broca's area, PMC, and the cerebellum) differentially process visual speech information, in the context of multisensory integration as well as during modality specific extraction of features to recover speech gesture information.

One potential confound that may exist for many studies that have investigated the brain regions involved with processing visual speech gesture information is the inability to distinguish whether the brain activity reflected processing of the visual gestural speech information or whether the brain activity reflected improved intelligibility that resulted from processes carried out elsewhere. Activity observed in many of the same brain regions thought to be involved with facilitative processing of visual speech information, including the PMC, Broca's area, Sylvian parietal temporal area Spt, IPL, and STG/S, have also been shown to be involved in increased intelligibility and comprehension (Callan et al., 2010; Londei et al., 2010). For studies of audio-visual speech processing this confound exists because in many cases the addition of visual speech gesture information improves intelligibility. A related confound is that it is often the case that these same brain regions (IFG, PMC, and cerebellum) involved with speech processing are also activated when task demands are high and require more working memory and attention (Jonides et al., 1998; Davachi et al., 2001; Sato et al., 2009; Alho et al., 2012). The activation of these regions may be related to task difficulty, greater attentional demand, and working memory (including internal rehearsal) that may be independent from specific processes involved with mapping between articulatory and auditory representations for speech perception. This increase in task demands occurs for most visual only speech tasks as well as for speech in noise tasks.

In this study the task was designed to control for both intelligibility and task difficulty by ensuring that performance using visual information alone was the same as that under the audio-visual conditions of interest. Specifically, we asked participants to identify vowels in visual and audio-visual speech stimuli. For this task, the visual information alone allowed for very high perceptual performance. Analyses focused on two regions of the PMC and the cerebellum, which have been previously shown to have mirror system properties and are thought to be involved in the instantiation of internal models (Callan et al., 2000a, 2004a, 2006a,b, 2010; Wilson et al., 2004; Skipper et al., 2007a). These regions are active during processing of visual speech information (Campbell et al., 2001; Bernstein et al., 2002; Nishitani and Hari, 2002; Olson et al., 2002; Callan et al., 2003, 2004a,b; Calvert and Campbell, 2003; Paulesu et al., 2003; Ojanen et al., 2005; Saito et al., 2005; Skipper et al., 2005, 2007b; Alho et al., 2012; Dubois et al., 2012; Mashal et al., 2012). One of these regions in the PMC

is more inferior and includes Broca's area and the PMvi. The other region is more superior and/or dorsal and has been referred to as PMvs and PMd.

It is rather uncontroversial that during the development of speech production, auditory-articulatory and orosensory-articulatory relationships must be established and encoded into internal models (Callan et al., 2000b; Tourville and Guenther, 2011; Guenther and Vladusich, 2012). Acoustic and orosensory signals are direct products of one's own articulation at are one goal of speech production. Likewise, internal models for visual aspects of speech (visual-auditory and visual-articulatory mappings) are learned by mapping features of speech gestures in the visual speech signal to the corresponding acoustics as well as to the articulations necessary to produce the corresponding deformation of the face. A primary goal of this study is to determine if the brain regions thought to instantiate internal models for speech (Broca's/PMvi, PMvs/PMd, IPL, Cerebellum) differ in their processing of audio-visual and visual only speech with respect to multisensory integration and modality specific extraction of articulatory speech gesture information (unimodal features in stimulation that specify phonemes). To accomplish this goal we identified the brain activity present during audio-visual and visual only speech processing. Given the results of previous experiments we hypothesized that both the PMvi/Broca's and PMvs/PMd would be active in both conditions. We further hypothesized the PMvi/Broca's area to be a site in which auditory and articulatory gesture information converge, and therefore activation in this area would show properties of multisensory enhancement. In contrast, a more prominent role for the PMvs/PMd may be the processing of modality specific speech gesture information. To determine which brain regions would show properties of multisensory enhancement we investigated differences in brain activity between audio-visual and audio only conditions at different signal-to-noise ratios. Based on the principle of inverse effectiveness (Wallace et al., 1992; Stein and Meredith, 1993) it was hypothesized that multisensory enhancement regions would show greater activity when unimodal audio stimuli had a lower signal-to-noise ratio.

METHODS

SUBJECTS

Sixteen 21–43 year-old (6 women and 10 men) right-handed subjects participated in this study. Eight subjects spoke English as their first language. The other eight subjects were native Japanese speakers who were proficient English speakers. The Japanese speakers all learned English beginning at 13 years of age or younger, and use English as their primary language at work and socially. Subjects gave written informed consent. The experimental procedures were approved by the ATR Human Subject Review Committee and were carried out in accordance with the principles expressed in the WMA Declaration of Helsinki.

PROCEDURE

Conditions

The experiment consisted of 10 conditions, however, only eight conditions were analyzed for this study. These eight conditions included: (1) an audiovisual condition (AV) where subjects saw a movie of the face articulating speech and heard the speaker utter a consonant-vowel-consonant (CVC) English monoslyllabic word with background audio noise (multispeaker babble) presented at three signal-to-noise ratios (−6, −10, and −14 dB; referred to as conditions AV6, AV10, AV14, respectively); (2) an audio only condition (A) where subjects saw a still face image while listening to the CVC with background audio noise at the same three signal-to-noise ratios (−6, −10, and −14 dB; referred to as conditions A6, A10, A14, respectively); (3) a visual only condition (VO) where subjects saw a movie of the face articulating speech, but without hearing the corresponding audio speech information or the audio noise; (4) and a baseline still face condition where subjects saw a still face but heard no audio. It should be note that in the same fMRI session subjects saw a still face with audio noise (SN) and a visual only condition with audio noise (VN) for a different study. The sound pressure level for the auditory stimuli was approximately 85–90 dB SPL. The stimuli were constructed such that the random segments of multispeaker babble noise were kept at a constant level and the speech signals were added to the babble noise at the specific signal to noise ratios (−6, −10, and −14 dB).

Protocol

The experiment consisted of a two-alternative forced choice task in which subjects identified by button press with their left thumb which vowel was present in the CVC English monoslyllabic word presented. In the baseline still face condition the subject randomly pushed one of the two buttons. The speech stimuli were spoken by a female native English speaker. Each presentation was 1 s in duration for all trials. For trials with visual speech this 1-s included facial motion before and after the audio speech signal for the word. The trial lasted approximately 3.9 s with ±200 ms of random jitter. The audio noise mixed with the speech signal consisted of an English multispeaker babble track (Audiotec, St. Louis, MO, USA). Multispeaker babble is known to be an effective and central masker of speech as its main energy is in the same range as the word stimuli (Wilson and Strouse, 2002). Three different runs were conducted each consisting of a separate vowel pair to be identified. The different vowel pairs consisted of /o-e/, /o-i/, and /o-^/ (^ as in gun). The stimuli were all common English words with pairs containing the same consonants (see **Table 1** for the list of stimuli). The left or right position of the button press for the /o/ response was counterbalanced across subjects and remained the same throughout the experiment for a single subject. Subjects were given practice trials before the experiment so they were familiar with the task and button response positions. Subjects were instructed to press the button to identify the vowel after presentation of each 1-s stimuli. The experimenter verbally instructed the subjects which button position was associated with each vowel before each run. There were seven different word pair stimuli for each vowel contrast (14 words for each vowel contrast). The same words were used for all the AV, A, and VO conditions. A blocked presentation design was implemented in which seven trials of the same condition were presented in succession for one block. The order of presentation of the various conditions was randomized. Subjects underwent three runs of fMRI scanning. Each run corresponded to a different vowel contrast to be identified, /o-e/, /o-i/, and /o-^/. The order of the vowel contrast runs

Table 1 | Stimulus word pairs used in experiment.

/o/-/e/	/o/-/i/	/o/-/^/
Cope–cape	Boat–beat	Coat–cut
Foam–fame	Gross–grease	Dome–dumb
Grove–grave	Load–lead	Phone–fun
Post–paste	Note–neat	Mode–mud
Prose–praise	Slope–sleep	Most–must
Toast–taste	Spoke–speak	Roast–rust
Woke–wake	Those–these	Tone–ton

was randomized across subjects. There were 20 blocks in each run. Each block lasted approximately 27.5 s. The 10 conditions were randomly presented in blocks of seven trials twice during each run. A block of seven trials for each condition was presented once before a block of trials of the same condition was presented the second time. In total there were 140 trials per run.

fMRI DATA COLLECTION AND PREPROCESSING

The visual speech signal was presented by means of a computer with specialized hardware and software that interfaced with a laser disk player containing the stimuli. The laser disk player was connected to the video projector. The video from the projector located outside of the MR room was directed to a mirror positioned inside of the head coil just above the subjects' eyes. The audio was presented via a sound file on the computer (premixed based on SNR) via MR-compatible headphones (Hitachi Advanced Systems' ceramic transducer headphones). The presentation of visual and audio signals using the computer hardware that controlled the laser disk ensured that there was no audio-visual asynchrony.

Brain imaging was conducted using a Shimadzu-Marconi's Magnex Eclipse 1.5T PD250 at the ATR Brain Activity Imaging Center. Functional T2* weighted images were acquired using a gradient echoplanar imaging sequence ($TR = 3.93$ s). An interleaved sequence was used consisting of 37 axial slices with a $4 \times 4 \times 4$ mm voxel resolution covering the cortex and cerebellum. Isotropic voxels were used to avoid possible distortion in realignment and normalization that occur with anisotropic voxels. For the scanner used in this study 3 mm voxels would have resulted in a longer than desired TR for each scan. Each run consisted of 140 scans. Images were preprocessed using programs within SPM8 (Wellcome Department of Cognitive Neurology, UCL). Differences in acquisition time between slices were accounted for, images were realigned and spatially normalized to MNI space ($3 \times 3 \times 3$ mm voxels) using the SPM template EPI image, and were smoothed using a $8 \times 8 \times 8$ mm FWHM Gaussian kernel. Regional brain activity for the various conditions was assessed using a general linear model employing a boxcar function convolved with a hemodynamic response function (global normalization and grand mean scaling were used to reduce artifacts). The baseline still face condition was implicitly modeled in the design. The nine other conditions were included in the SPM model. A fixed-effect analysis was first employed for all contrasts of interest for each subject. The contrast estimates of this analysis for each subject were used for random effects analysis. The contrasts of interest included the following: VO, AV (Combined Conditions AV6, AV10, AV14), VO-AV, AV-VO, multisensory enhancement (AV10-A10)-(AV6-A6) and (AV14-A14)-(AV10-A10). The threshold for significance was set at $p < 0.05$ using a False Discovery Rate FDR correction for multiple comparisons across the entire volume using a spatial extent threshold of 20 voxels. If no voxels were found to be significant using the FDR correction a threshold of $p < 0.001$ uncorrected with a spatial extent threshold of 20 voxels was used. Region of interest analyses were conducted using MNI coordinates for the PMv/IFG (−54, 6, 12), PMvs (−48, 0, 51), and the cerebellum (−12, −72, −45; 12, −72, −45) given in Callan et al. (2003) that were found to be important for audio visual processing. Bilateral coordinates in the cerebellum were used because studies have reported activity in both the left and right cerebellum in response to audio-visual speech (Callan et al., 2003; Saito et al., 2005; Skipper et al., 2005). Additionally, it is known that the cerebellum has predominantly crossed connections to the cortex such that the right hemisphere of the cerebellum projects to the language dominant left frontal areas including the PMC (Middleton and Strick, 1997; Schmahmann and Pandya, 1997). Small volume correction for multiple comparisons (pFWE < 0.05) were carried out using the seed voxels reported above within a sphere with a radius of 10 mm.

RESULTS

BEHAVIORAL RESULTS

Conditions showing better than chance performance

T-tests were used to determine which conditions showed performance that was significantly above chance on the two-alternative forced-choice vowel identification task (chance = 50%). There were 9 comparisons made altogether including the following: AV6, A6, AV10, A10, AV14, A14, AV All, A All, and VO. Bonferroni corrections for multiple comparisons were used to determine statistical significance at $p < 0.05$. Results of the analyses are presented in **Figure 1** and **Table 2**.

Audio-visual greater than audio only

A Two-Way analysis of Variance ANOVA was conducted over factors of Modality (with levels audio-visual and audio only) and SNR (with levels −6, −10, and −14 dB). Bonferroni corrections for multiple comparisons were used to determine statistical significance at $p < 0.05$ for planned ANOVA interaction and pairwise comparison analyses. In total there were seven planned analyses. The omnibus ANOVA indicated significant interaction between Modality and SNR, $F_{(2, 95)} = 7.1$, $p < 0.05$; and significant main effects of Modality (AV > A), $F_{(1, 95)} = 179.2$, $p < 0.05$, and SNR, $F_{(2, 95)} = 15.49$, $p < 0.05$. Planned pairwise comparisons (corrected for multiple comparisons) indicated statistically significant differences between the AV conditions and the A conditions (AV6-A6: $T = 5.79$, $p < 0.05$; AV10-A10: $T = 14.13$, $p < 0.05$, AV14-A14: $T = 14.2$, $p < 0.05$; AV > A: $T = 18.5$, $p < 0.05$; AV not significantly different from VO: $T = 0.69$; see **Figures 1**, **2**). The planned interaction analyses are given below.

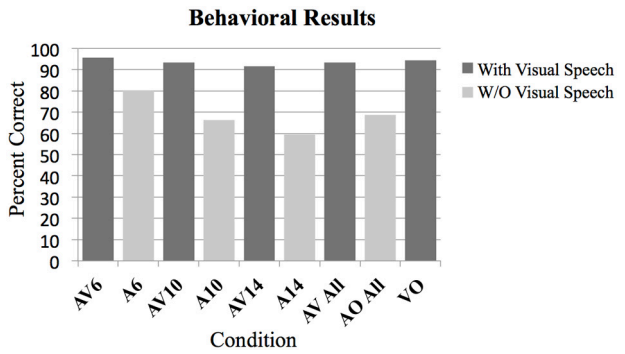


FIGURE 1 | Behavioral results as measured by percent correct on the two-alternative forced choice vowel identification task for the following conditions. Audio Visual AV6 (−6 dB SNR), Audio A6 (−6 dB SNR), AV10, A10, AV14, A14, All AV conditions combined, all A conditions combined, Video with noise VN, and Video only VO without noise. All contrasts were significantly greater than chance performance of 50% ($p < 0.01$).

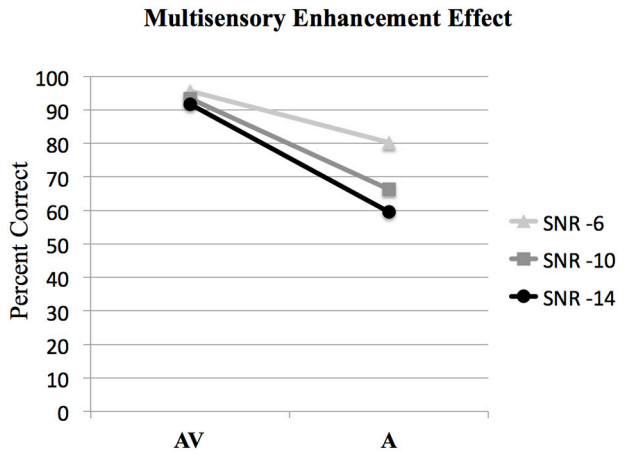


FIGURE 2 | Behavioral results showing the interaction of audio-visual enhancement at each of the signal-to-noise ratios SNRs. The interaction of (AV6-A6)-(AV10-A10) was statistically significant [$F_{(1, 15)} = 12.6$, $p < 0.005$]; however the interaction of (AV10-A10)-(AV14-A14) was not significant [$F_{(1, 15)} = 3.9$, $p > 0.05$].

Table 2 | *T*-Tests for conditions evaluating better than chance performance.

Condition	Mean %	*SE* %	*T*	Correct *p*
AV6	95.6	1.1	43.3	$p < 0.05$*
A6	80.3	2.9	10.4	$p < 0.05$*
AV10	93.4	2.0	21.7	$p < 0.05$*
A10	66.3	2.6	5.9	$p < 0.05$*
AV14	91.6	1.5	28.0	$p < 0.05$*
A14	59.5	2.7	3.5	$p > 0.05$
AV All	93.5	1.2	34.9	$p < 0.05$*
A All	68.7	2.3	8.0	$p < 0.05$*
VO	94.4	1.2	37.9	$p < 0.05$*

*Chance Performance was 50%. AV6, Audio-Visual −6 dB signal-to-noise ratio; A6 Audio Only −6 dB; AV10, Audio-Visual −10 dB; A10 Audio Only −10 dB; AV14, Audio-Visual −14 dB; A14 Audio Only −14 dB; SE, Standard Error; *significant using the Bonferroni correction for multiple comparisons.*

Multisensory enhancement effect

ANOVA was used to investigate interactions between AV and A conditions at different SNR levels to determine the presence of the multisensory enhancement effect. Bonferroni corrections for multiple comparisons were used to determine statistical significance at $p < 0.05$ for all analyses. The results of the analysis of the interaction between audio and visual conditions denoting the audio-visual enhancement effect are given in **Figure 2**. The interaction of (AV6-A6)-(AV10-A10) was statistically significant, [$F_{(1, 63)} = 8.2$, $p < 0.05$]. However, the interaction of (AV10-A10)-(AV14-A14) was not significant, $F_{(1, 63)} = 1.4$, $p > 0.05$ (see **Figure 2**).

Controlling for performance for conditions containing visual information

One of the goals of this experiment was to control for intelligibility and task difficulty across the different conditions containing visual information to determine which brain regions are involved with multisensory and visual speech gesture information processing. No significant difference was found between the combined audio-visual conditions AV and the VO condition using a lenient uncorrected threshold ($T = 0.69$, $p > 0.1$). This null effect is important for interpreting the fMRI results because ensuring that the perceptual performance across the conditions containing visual information did not differ was necessary (see **Figure 1**).

BRAIN IMAGING RESULTS

The random effect results of the fMRI analyses of the contrasts of interest are given in **Figures 3–8** and **Tables 3–7**. The brain activity rendered on the surface of the brain for the contrast of VO relative to baseline (still face plus button press) is given in **Figure 3**. Significant activity (pFDR < 0.05 corrected across entire volume; $T = 4.38$; see **Table 3** for detailed results) was present in left PMvi/Broca's area, left PMvs/PMd, left and right middle temporal visual motion processing area (MT/V5). The results of the ROI analysis showed significant activity ($p < 0.05$ corrected; see **Table 3**) in the left PMvi/Brocas area (MNI coordinate: −48, 9, 12), the left PMvs/PMd (MNI coordinate: −39, 3, 54). Significant activity (pFDR < 0.05 corrected across entire volume; $T = 3.28$) for the combined AV conditions was present in left and right PMvi/Broca's area, left PMvs/PMd, left and right STG/S, left MT/V5, and right cerebellum lobule VIIb (see **Figure 4** and **Table 4**). The results of the ROI analysis showed significant activity ($p < 0.05$ corrected; see **Table 4**) in the left PMvi/Broca's area (MNI coordinate: −51, 9, 9), the left PMvs/PMd (MNI coordinate: −48, 3, 42) and the right cerebellum lobule VIIb (MNI coordinate: 18, −72, −48). The conjunction of brain activity found to be active for both the combined AV conditions and the VO condition included the left PMvi/Broca's area, PMvs/PMd, and the left MT/V5 region (see **Figure 5**).

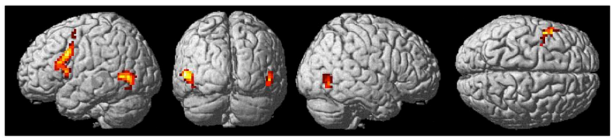


FIGURE 3 | Significant brain activity for the VO condition thresholded at pFDR < 0.05 corrected. Activity was present in the left PMvi/Broca's, left PMvs/PMd, and left and right MT/V5 visual motion processing area.

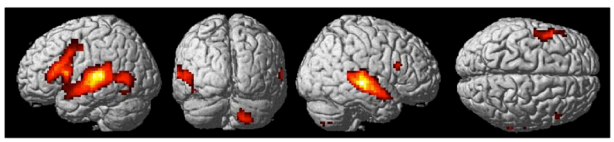


FIGURE 4 | Significant brain activity for the combined AV conditions thresholded at pFDR < 0.05 corrected. Activity was present in left and right PMvi/Broca's area, left PMvs/PMd, left and right STG/S including primary and secondary auditory cortex, left MT/V5 visual motion processing area, and the right cerebellum lobule VIIb.

FIGURE 5 | Brain activity that was significant for both (conjunction) the VO and the combined AV conditions thresholded at pFDR < 0.05 corrected. Activity was present in the left PMvi/Broca's, left PMvs/PMd, and left MT/V5 visual motion processing area.

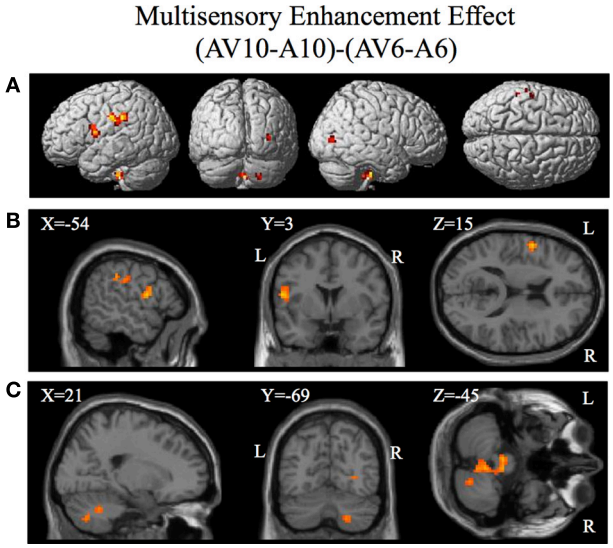


FIGURE 6 | Significant brain activity for the contrast that investigated the multisensory enhancement effect (AV10-A10)-(AV6-A6) thresholded at $p < 0.001$ uncorrected. Activity was present in left PMvi/Broca's area, left pre- and post-central gyrus, left inferior parietal cortex and suprmarginal gyrus, the right occipital lobe, the right cerebellum lobule VIIb and IX, and the left and right brain stem. **(A)** Activity rendered on the surface of the left, back, right, and top of the brain. **(B)** Section through brain taken at MNI coordinate −54, 3, 15 shows activity that was present in the PMvi and Broca's region. **(C)** Section through brain taken at MNI coordinate 21, −69, −45 shows activity that was present in cerebellum lobule VIIb. L, left side of brain; R, right side of brain.

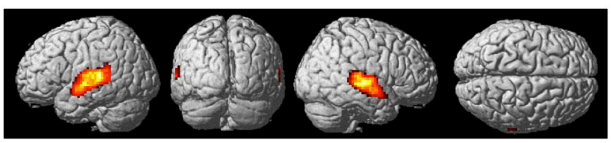


FIGURE 7 | Significant brain activity for the contrast of the combined AV conditions relative to the visual only VO condition thresholded at pFDR < 0.05 corrected. Activity was present in the left and right superior temporal gyrus/sulcus including primary and secondary auditory cortex.

Brain regions involved with the audio-visual enhancement effect across different signal-to-noise ratios were investigated using the contrast of (AV10-A10)-(AV6-A6) as well as the contrast of (AV14-A14)-(AV10-A10). The (AV10-A10)-(AV6-A6) contrast shows the degree of audio-visual enhancement as reflected in the behavioral results (see **Figure 2**) was greater when the signal-to-noise ratio was −10 dB compared to −6 dB. Significant activity was only found in the brain stem using the FDR correction for multiple comparisons, therefore the results are shown using a threshold of $p < 0.001$ ($T = 3.73$) uncorrected (see **Figure 6**). Active brain regions included the left PMvi/Broca's area, left pre-central gyrus (PreCG) Post central gyrus (PostCG), left inferior parietal cortex/supramarginal gyrus (IPC/SMG), right occipital lobe, the right cerebellar lobule VIIb and IX, and the left and right brain stem (see **Figure 6** and **Table 5**). The results of the ROI analysis showed significant activity ($p < 0.05$ corrected) in the left PMvi/Brocas area (MNI coordinate: −54, 3, 15), and the right cerebellum lobule VIIb (MNI coordinate: 21, −69, −45) (see **Table 5**). The behavioral results of the interaction of (AV14-A14)-(AV10-A10) did not show a significant multisensory enhancement effect (see **Figure 2**). Similarly, the results of the fMRI analysis for this contrast also did not reveal any significant activity ($p > 0.05$ uncorrected).

The contrasts investigating differences between the combined AV conditions and the VO condition are given in **Figures 7–8** and **Tables 6–7**. The contrast of AV vs. VO revealed significant activity (pFDR < 0.05 corrected across entire volume, $T = 3.48$) in only the STG/S region also encompassing primary and secondary auditory cortex (see **Figure 7** and **Table 6**). The results of the ROI analysis did not show any significant activity in the PMvi/Broca's, PMvs/PMd, or the cerebellum. The contrast of VO relative to the combined AV conditions did not show significant activity when using the FDR correction for multiple comparisons therefore the results are shown using a threshold of $p < 0.001$ uncorrected ($T = 3.73$; see **Figure 8**). Active brain regions include the left

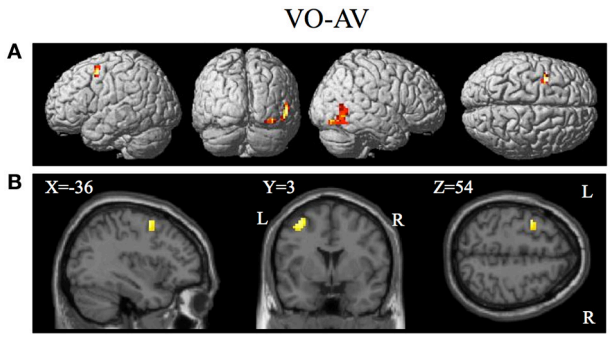


FIGURE 8 | Brain activity significantly active for the contrast of visual only VO relative to the combined AV conditions thresholded at $p < 0.001$ uncorrected. Activity was present in the left PMvs/PMd and the left MT/V5 visual motion processing area. **(A)** Activity rendered on the surface of the left, back, right, and top of the brain. **(B)** Section through brain taken at MNI coordinate −36, 3, 54 shows activity that was present in the PMvs/PMd region. L, left side of brain; R, right side of brain.

Table 3 | VO.

Brain region	MNI coordinates	*T*
PMvi/Broca's BA6, 44	−48, 12, 9	7.97
PMvs/PMd BA6	−39, 3, 54	4.70
MT/V5	−51, −69, 0	7.33
	54, −66, −3	6.07

Brain activity is thresholded using a false discovery rate FDR correction for multiple comparisons across the entire volume at pFDR < 0.05 for the Visual Only VO contrast. BA, Brodmann area; PMvi, Premotor ventral inferior; PMvs, Premotor ventral superior; PMd, Premotor dorsal; MT, Middle Temporal Gyrus; V5, Visual Area 5. Negative x MNI coordinates denote left hemisphere and positive x values denote right hemisphere activity.

PMvs/PMd, and the right MT/V5, and the right inferior occipital gyrus (see **Figure 8** and **Table 7**). The results of the ROI analysis (see **Table 7**) showed significant activity ($p < 0.05$ corrected) in the left PMvs/PMd (MNI coordinate: −39, 3, 54).

DISCUSSION

The purpose of this study was to determine if premotor regions, PMvi/Broca's and PMvs/PMd, as well as the cerebellum, demonstrate differential processing of multisensory (audio-visual) and unimodal (visual) speech gesture information. The primary finding was that the PMvi/Broca's area, the IPL, as well as the cerebellum showed properties of multisensory enhancement (see **Figure 6** and **Table 5**), while the PMvs/PMd showed greater unimodal visual only processing (see **Figure 8** and **Table 7**). It should be noted that activity in the speech motor areas, including the inferior frontal gyrus (including Broca's area) and a large portion of the PMC (including PMvi, PMvs, and PMd), was found for both the VO (see **Figure 3** and **Table 3**) and the AV (see **Figure 4** and **Table 4**) conditions. The activity in speech motor regions common to both of these conditions is shown by their conjunction in **Figure 5**.

Table 4 | AV.

Brain region	MNI coordinates *x, y, z*	*T*
PMvi/Broca's	−51, 9, 9	8.37
BA6 and 44	48, 18, 18	4.61
PMvs/PMd	−48, 3, 42	4.61
BA6		
STG/S	−51, −33, 9	12.08
BA 22, 41, 42	66, −24, 0	12.93
MT/V5	−51, −63, 6	5.78
CerbLob VIIb	18, −72, −48	5.5

Brain activity is thresholded using a false discovery rate FDR correction for multiple comparisons across the entire volume at pFDR < 0.05 for the combined (AV6, AV10, and AV14) audio visual AV contrast. BA, Brodmann area; PMvi, Premotor ventral inferior; PMvs, Premotor ventral superior; PMd, Premotor dorsal; STG/S, Superior temporal gyrus/sulcus; MT, Middle Temporal Gyrus; V5, Visual Area 5; CerbLob, Cerebellum Lobule. Negative x MNI coordinates denote left hemisphere and positive x values denote right hemisphere activity.

Table 5 | (AV10-A10)-(AV6-A6).

Brain region	MNI coordinates *x, y, z*	*T*
PMvi/Broca's BA6, 44	−54, 3, 15	5.2*
PreCG PostCG BA3, 4	−45, −18, 36	6.59
IPC/SMG BA40	−48, −36, 33	6.22
OccipLobe	33, −75, 6	4.91
CerbLob VIIb	21, −69, −45	4.38*
CerbLob IX	6, −51, −45	4.92
Brain stem	9, −30, −42	7.98**
	−6, −30, −42	5.75

*Brain activity is thresholded using p < 0.001 uncorrected, T = 3.73 for the multisensory enhancement contrast (AV10-A10)-(AV6-A6). BA, Brodmann area; PMvi, Premotor ventral inferior; PreCG, Pre-central gyrus; PostCG, Post-central gyrus; IPC, Inferior parietal cortex; SMG, Supramarginal Gyrus; OccipLobe, Occipital Lobe; CerbLob, Cerebellum Lobule. Negative x MNI coordinates denote left hemisphere and positive x values denote right hemisphere activity. *Denotes significant activity using a small volume correction for multiple comparisons with a 10 mm search radius (see Methods for seed voxel coordinates for ROIs). **Denotes significant (pFDR < 0.05) correction for multiple comparisons over the entire volume.*

It is often difficult to differentiate the brain networks that process the facial gestures that signal speech from the networks responsible for processing and integrating audio-visual speech stimuli because the intelligibility and task demands typically differ across conditions. Without controlling for these intelligibility differences, it is difficult to determine whether any increased brain activity reflects the processing of the visual and/or auditory features of speech, or is reflective of the level of intelligibility. As well, task difficulty can also confound the extent to which visual and audio-visual perception may show differential activity. This confound arises because activity in speech motor regions can be modulated by the degree of working memory and attention required for the speech task (Sato et al., 2009; Alho et al.,

Table 6 | AV-VO.

Brain region	MNI coordinates	*T*
STG/S	−45, −33, 6	13.2
BA22, 41, 42	57, −12, 3	11.23

Brain activity is thresholded using a false discovery rate FDR correction for multiple comparisons across the entire volume at pFDR < 0.05 for the combined audio-visual relative to the visual only VO contrast. BA, Brodmann area; STG/S, Superior Temproal Gyrus/Sulcus. Negative x MNI coordinates denote left hemisphere and positive x values denote right hemisphere activity.

Table 7 | VO-AV.

Brain region	MNI coordinates	*T*
PMvs/PMdBA6	−39, 3, 54	4.79*
MT/V5	51, −66, −9	5.07
IOG V4	36, −78, −12	5.69

*Brain activity is thresholded using p < 0.001 uncorrected, T = 3.73 for the visual only relative to the combined audio-visual contrast. BA, Brodmann area; PMvs, Premotor ventral superior; MT, Middle Temporal Gyrus; V5, Visual Area 5; IOG, Inferior Occipital Gyrus; V4, Visual area 4. Negative x MNI coordinates denote left hemisphere and positive x values denote right hemisphere activity. *Denotes significant activity using a small volume correction for multiple comparisons with a 10 mm search radius (see Methods for seed voxel coordinates for ROIs).*

2012). We controlled for intelligibility and task demands in this experiment by utilizing a vowel identification task in which the presentation of visual information alone allowed perceptual performance that was equally high as the performance observed for the audio-visual condition. Indeed, there were no significant differences in behavioral performance for the conditions containing visual information (see **Figure 1**). These results suggest that the intelligibility did not differ between conditions and that the task demands as far as general working memory and attention are concerned were essentially the same.

It was hypothesized that the PMvi/Broca's area is a site in which multisensory information (auditory, visual, orosensory) and speech gesture motor information are integrated and show properties of multimodal enhancement (Wallace et al., 1992; Stein and Meredith, 1993; Callan et al., 2003). The brain imaging results (see **Figure 6**) of the (AV10-A10)-(AV6-A6) contrast showed activity related to the audio-visual enhancement effect (see **Figure 2**) when the signal-to-noise ratio of the audio signal was reduced. Of particular interest is activity denoting multisensory enhancement in the left hemisphere PMvi/Broca's, pre- and post-central gyrus, the IPC/SMG and the right cerebeullum lobule VIIb. These areas are all thought to be involved with forward and inverse internal models used to facilitate speech perception (Callan et al., 2004a; Rauschecker, 2011). Although these properties of multisensory enhancement were found in the PMvi/Broca's area it is not the case that this area was more strongly activated by the audio-visual stimuli than it was by the visual only stimuli in this study. The contrast of AV-V (see **Figure 7** and **Table 6**) only shows activity in the STG/S and no significant activity even in the ROI analysis within PMvi/Broca's area. It is unclear why multisensory enhancement was not found in the STG/S, considering that multisensory enhancement has been observed in this area in other studies (Calvert et al., 2000; Callan et al., 2001, 2003, 2004b). It may not be too surprising that the brain imaging contrast between (AV14-A14)-(AV10-A10) did not show any significant brain activity given that the behavioral visual enhancement effect was also not significant (see **Figure 2**). One potential reason for the lack of an enhancement effect for this contrast may be that the audio signal was so low that there was not enough auditory information available to integrate with the visual information. This hypothesis is supported by the fact that the A14 condition did not significantly differ from chance performance, when corrections were made for multiple comparisons (see **Figure 2** and **Table 2**).

We hypothesized that the PMvs/PMd region is involved with mapping unimodal aspects of sensory information onto speech articulatory gestures. The contrast of the visual only relative to the combined audio-visual conditions V-AV (see **Figure 8**, **Table 7**) showed activity in the left PMvs/PMd and the left MT/V5. The finding of differential activity in visual motion processing area MT/V5 is consistent with the assertion that a greater reliance on information in visual speech motion features is utilized when auditory information is not present. It is important to note that this activity is not a result of differences in task difficulty or intelligibility as these were the same between visual only V and audio-visual AV conditions.

The results of this study are consistent with the hypothesis that overlapping processes are carried out by PMvi/Broca's region and the PMvs/PMd region but that processing in these areas differ in the degree to which they process multisensory and unimodal stimuli. Within the context of an internal model based approach we propose that the nervous system relies to a greater degree on visual-articulatory based mappings when stimulus driven auditory-articulatory based mappings are not present. One could further conjecture that the PMvi/Broca's region may be more influenced by the ventral stream (what pathway) and the PMvs/PMd may be more influenced by the dorsal stream (where/how pathway). This is consistent with the model proposed by (Rauschecker and Scott, 2009; Rauschecker, 2011) in which the antero-ventral stream includes Broca's area PMv and the postero-dorsal stream includes the PMd. Multiple fiber tracts (Friederici, 2009) from superior temporal areas to IFG and PMC give support to the possibility of both antero-ventral and postero-dorsal streams including frontal speech regions. The inclusion of frontal speech areas in both the antero-ventral and postero-dorsal streams is in contrast to the model proposed by (Hickok and Poeppel, 2000, 2004, 2007) in which it is proposed that frontal speech areas (Broca's/PMvi; PMvs/PMd) are all thought to be within the postero-dorsal stream.

ACKNOWLEDGMENTS

This research was supported by the National Institute of Information and Communications Technology and by KAKENHI, Grant-in-Aid for Scientific Research(C) (21500321).

REFERENCES

Alho, J., Sato, M., Sams, M., Schwartz, J., Tiitinen, H., and Jaaskelainen, I. (2012). Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. *Neuroimage* 60, 1937–1946. doi: 10.1016/j.neuroimage.2012.02.011

Bernstein, L., Auer, E., Moore, J., Ponton, C., Don, M., and Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport* 13, 311–315. doi: 10.1097/00001756-200203040-00013

Callan, A., Callan, D., Tajima, K., and Akahane-Yamada, R. (2006a). Neural processes involved with perception of non-native durational contrasts. *Neuroreport* 17, 1353–1357. doi: 10.1097/01.wnr.0000224774.66904.29

Callan, D., Callan, A., Gamez, M., Sato, M., and Kawato, M. (2010). Premotor cortex mediates perceptual performance. *Neuroimage* 51, 844–858. doi: 10.1016/j.neuroimage.2010.02.027

Callan, D., Callan, A., Honda, K., and Masaki, S. (2000a). Single-sweep EEG analysis of neural processes underlying perception and production of vowels. *Cogn. Brain Res.* 10, 173–176. doi: 10.1016/S0926-6410(00)00025-2

Callan, D. E., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: a singlesweep EEG case study. *Cogn. Brain Res.* 10, 349–353. doi: 10.1016/S0926-6410(00)00054-9

Callan, D., Jones, J., Callan, A., and Akahane-Yamada, R. (2004a). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *Neuroimage* 22, 1182–1194. doi: 10.1016/j.neuroimage.2004.03.006

Callan, D., Jones, J., Munhall, K., Callan, A., Kroos, C., and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport* 14, 2213–2218. doi: 10.1097/00001756-200312020-00016

Callan, D., Jones, J., Munhall, K., Kroos, C., Callan, A., and Vatikiotis-Bateson, E. (2004b). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J. Cogn. Neurosci.* 16, 805–816. doi: 10.1162/089892904970771

Callan, D., Kawato, M., Parsons, L., and Turner, R. (2007). Speech and song: The role of the cerebellum. *Cerebellum* 6, 321–327. doi: 10.1080/14734220601187733

Callan, D., Kent, R., Guenther, F., and Vorperian, H. (2000b). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *J. Speech Lang. Hear. Res.* 43, 721–736.

Callan, D., and Manto, M. (2013). "Cerebellar control of speech and song," in *Handbook of the Cerebellum and Cerebellar Disorders,* eds M. Manto, D. Gruol, J. Schmahmann, N. Koibuchi, and F. Rossi (New York, NY: Springer).

Callan, D., Tsytsarev, V., Hanakawa, T., Callan, A., Katsuhara, M., Fukuyama, H., et al. (2006b). Song and speech: brain regions involved with perception and covert production. *Neuroimage* 31, 1327–1342. doi: 10.1016/j.neuroimage.2006.01.036

Calvert, G. A., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* 15, 57–70. doi: 10.1162/089892903321107828

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10, 649–657. doi: 10.1016/S0960-9822(00)00513-3

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G. A., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cogn. Brain Res.* 12, 245–264. doi: 10.1016/S0926-6410(01)00054-4

Davachi, L., Maril, A., and Wagner, A. D. (2001). When keeping in mind supports later bringing to mind: neural markers of phonologi- cal rehearsal predict subsequent remembering. *J. Cogn. Neurosci.* 13, 1059–1070. doi: 10.1162/089892901753294356

Dubois, C., Otzenberger, H., Gounout, D., Sock, R., and Metz-Lutz, M. N. (2012). Visemic processing in audiovisual discrimination of natural speech: a simultaneous fMRI-EEG study. *Neuropsychologia* 50, 1316–1326. doi: 10.1016/j.neuropsychologia.2012.02.016

Friederici, A. (2009). Pathways to language: fiber tracts in the human brain. *Trends Cogn. Sci.* 13, 175–181. doi: 10.1016/j.tics.2009.01.001

Grant, K. W., and Braida, L. D. (1991). Evaluating the articulation index for audiovisual input. *J. Acoust. Soc. Am.* 89, 2952–2960. doi: 10.1121/1.400733

Guenther, F., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neurolinguist.* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Hickok, G., and Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends Cogn. Sci.* 4, 131–128. doi: 10.1016/S1364-6613(00)01463-7

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Iacoboni, M. (2008). The role of premotor cortex in speech perception: evidence from fMRI and rTMS. *J. Physiol.* 102, 31–34. doi: 10.1016/j.jphysparis.2008.03.003

Iacoboni, M., and Wilson, S. (2006). Beyond a single area: motor control and language within a neural architecture encompassing Broca's area. *Cortex* 42, 503–506. doi: 10.1016/S0010-9452(08)70387-3

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Putz, B., et al. (2000). Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* 403, 192–195. doi: 10.1038/35003194

Jonides, J., Schumacher, E. H., Smith, E. E., Koeppe, R. A., Awh, E., Reu-ter-Lorenz, P. A., et al. (1998). The role of pa- rietal cortex in verbal working memory. *J. Neurosci.* 18, 5026–5034.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* 9, 718–727. doi: 10.1038/35003194

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (1967). Perception of speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Liberman, A., and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Liberman, A., and Whalen, D. (2000). On the relation of speech to language. *Trends Cogn. Sci.* 4, 187–196. doi: 10.1016/S1364-6613(00)01471-6

Londei, A., D' Ausilio, A., Basso, D., Sestieri, C., Del Gratta, C., Romani, G., et al. (2010). Sensory-motor brain network connectivity for speech comprehension. *Hum. Brain Mapp.* 31, 567–580. doi: 10.1002/hbm.20888

Mashal, N., Solodkin, A., Dick, A., Chen, E., and Small, S. (2012). A network model of observation and imitation of speech. *Front. Psychology*. 3:84. doi: 10.3389/fpsyg.2012.00084

Meister, I., Wilson, S., Deblieck, C., Wu, A., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Middleton, F. A., and Strick, P. (1997). Cerebellar output channels. *Int. Rev. Neurobiol.* 41, 61–82. doi: 10.1016/S0074-7742(08)60347-5

Nishitani, N., and Hari, R. (2002). Viewing lip forms: cortical dynamics. *Neuron* 36, 1211–1220. doi: 10.1016/S0896-6273(02)01089-9

Nishitani, N., Schurmann, M., Amunts, K., and Hari, R. (2005). Broca's region: from action to language. *Physiology* 20, 60–69. doi: 10.1152/physiol.00043.2004

Ojanen, V., Mottonen, R., Pekkol, J., Jaaskelainen, I., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Brocas area. *Neuroimage* 25, 333–338. doi: 10.1016/j.neuroimage.2004.12.001

Olson, I. R., Gatenby, J. G., and Gore, J. C. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Cogn. Brain Res.* 14, 129–138. doi: 10.1016/S0926-6410(02)00067-8

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., et al. (2003). A functional-anatomical model for lip-reading. *J. Neurophysiol.* 90, 2005–2013. doi: 10.1038/35003194

Poeppel, D., Idsardi, W. J., and van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1071–1086. doi: 10.1098/rstb.2007.2160

Rauschecker, J. (2011). An expanded role fore the dorsal auditory pathway in sensorimotor control and integration. *Hear. Res.* 271, 16–25. doi: 10.1016/j.heares.2010.09.001

Rauschecker, J., and Scott, S. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Saito, D., Yoshimura, K., Kochiyama, T., Okada, T., Honda, M., and Sadato, N. (2005). Cross-modal binding and activated attentional networks during audio-visual speech integration: a functional MRI study. *Cereb. Cortex* 15, 1750–1760. doi: 10.1093/cercor/bhi052

Sato, M., Tremblay, P., and Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7. doi: 10.1016/j.bandl.2009.03.002

Schmahmann, J., and Pandya, D. N. (1997). The cerebrocerebellar system. *Int. Rev. Neurobiol.* 41, 31–60. doi: 10.1016/S0074-7742(08)60346-3

Schwartz, J., Basirat, A., Menard, L., and Sato, M. (2012). The perception-for-action-control theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguist.* 25, 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual perception examined by fMRI and PET. *Neurosci. Res.* 47, 277–287. doi: 10.1016/S0168-0102(03)00214-1

Skipper, J., Goldin-Meadow, S., Nusbaum, H., and Small, S. (2007a). Speech-associated gestures, Broca's area, and the human mirror system. *Brain Lang.* 101, 260–277. doi: 10.1016/j.bandl.2007.02.008

Skipper, J., van Wassenhove, V., Nusbaum, H., and Small, S. (2007b). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Skipper, S., Nusbaum, H., and Small, S. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89. doi: 10.1016/j.neuroimage.2004.11.006

Stein, B., and Meredith, M. (1993). *The Merging of the Senses*. Cambridge: MIT Press.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Accoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Tourville, J. Guenther, F. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. *Exp. Brain Res.* 91, 484–488. doi: 10.1007/BF00227844

Wilson, R., and Strouse, A. (2002). Northwestern University auditory test no. 6 in multi-talker babble: a preliminary report. *J. Rehabil. Res. Dev.* 39, 105–114.

Wilson, S., and Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage* 33, 316–325. doi: 10.1016/j.neuroimage.2006.05.032

Wilson, S., Saygin, A., Sereno, M., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 January 2014; accepted: 14 April 2014; published online: 05 May 2014.
Citation: Callan DE, Jones JA and Callan A (2014) Multisensory and modality specific processing of visual speech in different regions of the premotor cortex. Front. Psychol. ***5****:389. doi: 10.3389/fpsyg.2014.00389*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 03 June 2014
doi: 10.3389/fpsyg.2014.00427

Speech is not special... again

Kathy M. Carbonell and Andrew J. Lotto*

Department of Speech, Language and Hearing Sciences, University of Arizona, Tucson, AZ, USA
**Correspondence: alotto@email.arizona.edu*

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Jean Vroomen, University of Tilburg, Netherlands

Keywords: sensorimotor effects on perception, multisensory integration, speech perception, auditory processing, Motor Theory

THE "SPECIALNESS" OF SPEECH

As is apparent from reading the first line of nearly any research or review article on speech, the task of perceiving speech sounds is complex and the ease with which humans acquire, produce and perceive these sounds is remarkable. Despite the growing appreciation for the complexity of the perception of music, speech perception remains the most amazing and poorly understood auditory (and, if we may be so bold, perceptual) accomplishments of humans. Over the years, there has been considerable debate on whether this achievement is the result of general perceptual/cognitive mechanisms or "special" processes dedicated to the mapping of speech acoustics to linguistic representations (for reviews see Trout, 2001; Diehl et al., 2004). The most familiar proposal of the "specialness" of speech perception is the various incarnations of the Motor Theory of speech proposed by Liberman et al. (1967; Liberman and Mattingly, 1985, 1989). Given the status of research into audition in the 1950s and 1960s, it is not surprising that speech appeared to require processing not available in "normal" hearing. Much of the work at the time used relatively simple tones and noises to get at the basic psychoacoustics underlying the perception of pitch and loudness (though some researchers like Harvey Fletcher were also working on some basics of speech perception, Fletcher and Galt, 1950; Allen, 1996). Liberman and his collaborators discovered that the discrimination of acoustic changes in speech sounds did not look like the psychoacoustic measures of discrimination for pitch and loudness. Instead of following a Weber or Fechner law, the discrimination function had a peak near the categorization boundary between contrasting phonemes—a pattern of perceptual results that is referred to as Categorical Perception (Liberman et al., 1957). In addition, the acoustic cues to phonemic identity were not readily apparent with similar spectral patterns resulting in different phonemic percepts and acoustically disparate patterns resulting in identical phonemic percepts—the problem of "lack of invariance" (e.g., Liberman et al., 1952). The perception of these varying acoustic patterns was highly context-sensitive to preceding and following phonetic content in ways that appeared specific to the communicative constraints of speech and not applicable to the perception of other sounds—as in demonstrations of perceptual compensation for coarticulation, speaking rate normalization and talker normalization (e.g., Ladefoged and Broadbent, 1957; Miller and Liberman, 1979; Mann, 1980).

One major source of evidence in favor of a Motor Theory account of speech perception is that information about a speaker's production (anatomy or kinematics) from non-auditory sources can affect phonetic perception. The famed McGurk effect (McGurk and MacDonald, 1976), in which visual presentation of a talker can alter the auditory phonetic percept, is taken as evidence that listeners are integrating information about production from this secondary source. Fowler and Deckle (1991) have demonstrated a similar effect using haptic information gathered by touching the speaker's face (see also Sato et al., 2010). Gick and Derrick (2009) reported that perception of consonant—vowel tokens in noise are biased toward voiceless stops (e.g., /pa/) when they are accompanied by a small burst of air on the skin of the listener, which could be interpreted as the aspiration that would more likely accompany the release of a voiceless stop.

In addition, there have been several studies that have demonstrated that manipulations of the *listener's* articulators can affect perception, which are supportive of the Motor Theory proposal that the mechanisms of production underlie the perception of speech. For example, Ito et al. (2009) obtained shifts in phoneme categorization resulting from external manipulation of the skin around the *listener's* mouth in ways that would correspond to the deformations typical of producing these speech sounds (see also Yeung and Werker, 2013 for a similar demonstration with infants). Recently, Mochida et al. (2013) found that the ability to categorize consonants can be influenced by the simultaneous silent production of these consonants. Typically, these studies are proffered as evidence for a direct role of speech motor processing in speech perception.

Independent of this proposed motor basis of perception, others have suggested the existence of a special speech or phonetic mode of perception based on evidence of neural and behavioral responses to the same stimuli being modulated by whether or not the listener believes the signal to be speech or non-speech (e.g., Tomiak et al., 1987; Vroomen and Baart, 2009; Stekelenburg and Vroomen, 2012).

THE "GENERALITY" OF SPEECH

Since the early work by Liberman and colleagues and the development of the Motor Theory, there has been a growing appreciation for the power of perceptual learning and the context-sensitive nature of auditory processing. Once one begins to study more complex sounds and perceptual behaviors, the distinction between speech

and non-speech processing becomes less clear. So, for example, we now have many examples of non-speech sound categories that demonstrate the characteristics of Categorical Perception (Cutting et al., 1976; Harnad, 1990; Mirman et al., 2004). It also appears that general auditory learning mechanisms are capable of dealing with the lack of invariance problem in formation of categories. Birds can learn speech consonant categories with no obvious acoustic invariant cue (Kluender et al., 1987) and human listeners can readily learn non-speech categories that are similarly structured (Wade and Holt, 2005). Finally, non-speech analogs have been created that result in the same types of context effects earlier witnessed for speech categorization, such as "perceptual compensation for coarticulation" (Lotto and Kluender, 1998; Holt et al., 2000), "speaking rate normalization" (Pisoni et al., 1983; Diehl and Walsh, 1989) and "talker normalization" (Watkins and Makin, 1994; Holt, 2005; Sjerps et al., 2011; Laing et al., 2012).

These findings with non-speech and animal perception of speech sounds (along with many others) call into question the strict dichotomy of speech and general auditory processing (Schouten, 1980). The lack of a clear distinction extends to the famed McGurk effect, which has been successfully modeled using general models of perception (e.g., Massaro, 1998). Stephens and Holt (2010) demonstrated that human adults can learn correlations between features of speech and arbitrary dynamic visual cues that are not related to the gestures of human vocal tracts. Participants in their experiments learned to associate the movements of dials and lighted bars on an animated "robot" display to stimuli varying in vowels and voiced consonant and could use this information to enhance intelligibility in noise. These types of novel mappings demonstrate the effectiveness of perceptual learning even across modalities (though perhaps not leading to as strong of an integration of information as may occur for natural covariations).

THE IMPORTANCE OF RESEARCH INTO MULTISENSORY INTERACTIONS IN SPEECH PERCEPTION

The growth in empirical research into the integration of multisensory information in speech acquisition and perception is a welcome development because it is a recognition that speech is not perceived within a vacuum. Too often, speech perception research has been conducted in an isolated reductionist vein that has made the human accomplishments in speech communication seem almost miraculous. The important realization at the heart of Lindblom's (1990, 1996) Hypo and Hyper Speech Theory is that much of the troubling acoustic variability in speech is actually a result of the changing demands of conversation between two people and the needs for informational precision due to the communication context. When one fails to study speech within a full communication context, this structured variability becomes noise. The isolation of speech research from a communication context has also made it difficult to connect the vast work in phonemic perception with more practical clinical issues in hearing loss and speech pathology. As Weismer and Martin (1992) point out, the concept of intelligibility must include both the speaker and the listener—that is, intelligibility is a measure of the entire communication setting and not just the acoustics of the speaker (see also, Liss, 2007).

The investigation of multisensory integration in speech perception is a step in the direction of attempting to understand the entire communication setting and all of the available information that results in an intelligible message. Some of the well-known findings from an auditory-isolated experiment may in fact be misleading when looked at in this broader context. For example, a highly cited finding is that 9-month-old infants from English-speaking households fail to discriminate a non-native Hindi contrast (Werker and Tees, 1984), which is taken as evidence that they are now perceptually tuned to their native language. However, Yeung and Werker (2009) obtained discrimination for infants in this group when the contrasting sounds were paired consistently with visual novel objects—a situation which mimics more realistically the communication setting of language learning. MacKenzie et al. (2013) in one experiment demonstrated an apparent unwillingness of 12-month-olds to associate novel auditory words with visual objects when the words are not phonotactically acceptable in their native language. However, the infants show far more flexibility in "acceptable" words when the task is preceded by a word-object association game with familiar word-objects. In each of these examples, the presumed perceptual tuning for language becomes less strict once the information available to the infant about the task is expanded. These experiments are stark reminders that speech acquisition and perception occurs in a larger perceptual/cognitive framework. Such results may also extend to adults learning to categorize speech sounds. Lim and Holt (2011) obtained significant increases in categorization performance for Japanese-speaking adults learning the non-native English /l/-/r/ distinction utilizing a video game paradigm. In this game, the categories were associated with different visual creatures that were either "friends" or "enemies" requiring different actions. The implicit mapping of auditory categories to functional dynamic visual objects may account for some of the success of this training.

A CAUTIONARY NOTE

Whereas the section above provides just a few of the many benefits of studying multisensory integration in speech, one must be cautious not to repeat the history of the field by proposing special mechanisms of phenomena for speech perception without thoroughly investigating what processes are available for general perception. The perception of all sound events is almost certainly intrinsically multisensory. Experimental designs that reduce sound event perception to audition run the risk of changing the task demands for the perceiver (as seen above in the examples for speech discrimination in infants).

There are many examples of sound perception being influenced by non-auditory information. Detection of low-intensity sounds is enhanced when paired with a task-irrelevant light stimulus (Lovelace et al., 2003; Odgaard et al., 2004). Saldaña and Rosenblum (1993) reported that when listeners were presented a visual image of a cello either being plucked or bowed, it strongly influenced their auditory judgment of whether the cello was being plucked or bowed. The perceived loudness of tones can be influenced by synchronous tactile information (Schürmann et al., 2004; Gillmeister and Eimer, 2007).

In addition, sensori-motor interactions can be found in music perception (Maes et al., 2013). We should be very cautious in proposing multimodal or sensorimotor interactions that are "special" to speech. It is quite possible that new integrations between senses will be observed using the well-learned complex stimuli of speech sounds (or musical sounds) as opposed to simple noises and tones and unexperienced complex signals. These novel findings should be taken as opportunities to learn general principles of perception, action and cognition as opposed to assigning them special status and missing these opportunities.

Postulating a special speech perception mode or module is a strong theoretical position not to be taken lightly. One must describe how the processes brought to bear in the perception of speech sounds are fundamentally different from those responsible for other forms of complex audition. Speech sounds are "special" in the sense that they are over-learned categories that play a functional role in a larger hierarchical linguistic system. But these attributes on their own do not necessitate the proposal of inherently different processing mechanisms. In the end, speech sounds and the perception/categorization of these sounds is not likely to require special processing. The "specialness" of these sounds comes from being a part of the complex act of communicating. It is the act of communicating that clearly requires integration of the senses and the cooperation of perception and action. We must be wary that speech sound perception ("is this a "ba" or a"da") isolated from the full act of communication is unnatural even when bringing to bear information from other sense modalities. The small and context-specific sensorimotor and multisensory effects we can uncover in this artificial task (Hickok et al., 2009) may not provide much insight into the real act of communication with speech.

REFERENCES

Allen, J. B. (1996). Harvey Fletcher's role in the creation of communication acoustics. *J. Acoust. Soc. Am.* 99, 1825–1839. doi: 10.1121/1.415364

Cutting, J. E., Rosner, B. S., and Foard, C. F. (1976). Perceptual categories for musiclike sounds: implications for theories of speech perception. *Q. J. Exp. Psychol.* 28, 361–378. doi: 10.1080/14640747608400563

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.* 55, 149–179. doi: 10.1146/annurev.psych.55.090902.142028

Diehl, R. L., and Walsh, M. A. (1989). An auditory basis for the stimulus−length effect in the perception of stops and glides. *J. Acoust. Soc. Am.* 85, 2154–2164. doi: 10.1121/1.397864

Fletcher, H., and Galt, R. H. (1950). The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.* 22, 89–151. doi: 10.1121/1.1906605

Fowler, C., and Deckle, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 816–828. doi: 10.1037/0096-1523.17.3.816

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462, 502–504. doi: 10.1038/nature08572

Gillmeister, H., and Eimer, M. (2007). Tactile enhancement of auditory detection and perceived loudness. *Brain Res.* 1160, 58–68. doi: 10.1016/j.brainres.2007.03.041

Harnad, S. R. (Ed). (1990). *Categorical Perception: the Groundwork of Cognition*. New York, NY: Cambridge University Press.

Hickok, G., Holt, L. L., and Lotto, A. J. (2009). Response to Wilson: what does motor cortex contribute to speech perception? *Trends Cogn. Sci.* 13, 330–331. doi: 10.1016/j.tics.2009.05.002

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychol. Sci.* 16, 305–312. doi: 10.1111/j.0956-7976.2005.01532.x

Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *J. Acoust. Soc. Am.* 108, 710–722. doi: 10.1121/1.429604

Ito, T., Tiede, M., and Ostry, D. J. (2009). Somatosensory function in speech perception. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1245–1248. doi: 10.1073/pnas.0810063106

Kluender, K. R., Diehl, R. L., and Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science* 237, 1195–1197. doi: 10.1126/science.3629235

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98–104. doi: 10.1121/1.1908694

Laing, E. J. C., Liu, R., Lotto, A. J., and Holt, L. L. (2012). Tuned with a tune: talker normalization via general auditory processes. *Front. Psychol.* 3:203. doi: 10.3389/fpsyg.2012.00203

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Liberman, A. M., Delattre, P., and Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am. J. Psychol.* 65, 497–516. doi: 10.2307/1418032

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358. doi: 10.1037/h0044417

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Liberman, A. M., and Mattingly, I. G. (1989). A specialization for speech perception. *Science* 243, 489–494. doi: 10.1126/science.2643163

Lim, S.-J., and Holt, L. L. (2011). Learning foreign sounds in an alien world: video game training improves non-native speech categorization. *Cogn. Sci.* 35, 1390–1405. doi: 10.1111/j.1551-6709.2011.01192.x

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the HandH theory," in *Speech Production and Speech Modelling*, eds W. Hardcastle and M. Kluwer (Netherlands: Springer), 403–439. doi: 10.1007/978-94-009-2037-8_16

Lindblom, B. (1996). Role of articulation in speech perception: clues from production. *J. Acoust. Soc. Am.* 99, 1683–1692. doi: 10.1121/1.414691

Liss, J. M. (2007). "Perception of dysarthric speech," in *Motor Speech Disorders: Essays for Ray Kent*, ed G. Weismer (San Diego, CA: Plural Publishing), 187–219.

Lotto, A. J., and Kluender, K. R. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Percept. Psychophys.* 60, 602–619. doi: 10.3758/BF03206049

Lovelace, C. T., Stein, B. E., and Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Brain Res. Cogn. Brain Res.* 17, 447–453. doi: 10.1016/S0926-6410(03)00160-5

MacKenzie, H. K., Graham, S. A., Curtin, S., and Archer, S. L. (2013). The flexibility of 12-month-olds' preferences for phonologically appropriate object labels. *Dev. Psychol.* 50, 422–430. doi: 10.1037/a0033524

Maes, P. J., Leman, M., Palmer, C., and Wanderley, M. M. (2013). Action-based effects on music perception. *Front. Psychol.* 4:1008. doi: 10.3389/fpsyg.2013.01008

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Percept. Psychophys.* 28, 407–412. doi: 10.3758/BF03204884

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Vol. 1. Cambridge, MA: MIT Press.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Miller, J. L., and Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Percept. Psychophys.* 25, 457–465. doi: 10.3758/BF03213823

Mirman, D., Holt, L. L., and McClelland, J. L. (2004). Categorization and discrimination of nonspeech sounds: differences between steady-state and rapidly-changing acoustic cues. *J. Acoust. Soc. Am.* 116, 1198–1207. doi: 10.1121/1.1766020

Mochida, T., Kimura, T., Hiroya, S., Kitagawa, N., Gomi, H., and Kondo, T. (2013). Speech misperception: speaking and seeing interfere differently with hearing. *PLoS ONE* 8:e68619. doi: 10.1371/journal.pone.0068619

Odgaard, E. C., Arieh, Y., and Marks, L. E. (2004). Brighter noise: sensory enhancement of perceived loudness by concurrent visual stimulation. *Cogn. Affect. Behav. Neurosci.* 4, 127–132. doi: 10.3758/CABN.4.2.127

Pisoni, D. B., Carrell, T. D., and Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Percept. Psychophys.* 34, 314–322. doi: 10.3758/BF03203043

Saldaña, H. M., and Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Percept. Psychophys.* 54, 406–416.

Sato, M., Cavé, C., Ménard, L., and Brasseur, A. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia* 48, 3683–3686. doi: 10.1016/j.neuropsychologia.2010.08.017

Schouten, M. E. H. (1980). The case against a speech mode of perception. *Acta Psychol.* 44, 71–98. doi: 10.1016/0001-6918(80)90077-3

Schürmann, M., Caetano, G., Jousmäki, V., and Hari, R. (2004). Hands help hearing: facilitatory audiotactile interaction at low sound-intensity levels. *J. Acoust. Soc. Am.* 115, 830–832. doi: 10.1121/1.1639909

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Atten. Percept. Psychophys.* 73, 1195–1215. doi: 10.3758/s13414-011-0096-8

Stekelenburg, J. J., and Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. *Neuropsychologia* 50, 1425–1431. doi: 10.1016/j.neuropsychologia.2012.02.027

Stephens, J. D. W., and Holt, L. L. (2010). Learning novel artificial visual cues for use in speech identification. *J. Acoust. Soc. Am.* 128, 2138–2149. doi: 10.1121/1.3479537

Tomiak, G. R., Mullennix, J. W., and Sawusch, J. R. (1987). Integral processing of phonemes: evidence for a phonetic mode of perception. *J. Acoust. Soc. Am.* 81, 755–764. doi: 10.1121/1.394844

Trout, J. D. (2001). The biological basis of speech: what to infer from talking to the animals. *Psychol. Rev.* 108, 523–549. doi: 10.1037/0033-295X.108.3.523

Vroomen, J., and Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition* 110, 254–259 doi: 10.1016/j.cognition.2008.10.015

Wade, T., and Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J. Acoust. Soc. Am.* 118, 2618–2633. doi: 10.1121/1.2011156

Watkins, A. J., and Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral—envelope distortion. *J. Acoust. Soc. Am.* 96, 1263–1282. doi: 10.1121/1.410275

Weismer, G., and Martin, R. (1992). "Acoustic and perceptual approaches to the study of intelligibility," in *Intelligibility in Speech Disorders: Theory, Measurement and Management*, ed R. D. Kent (Amsterdam: John Benjamins), 67–118.

Werker, J. F., and Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* 7, 49–63. doi: 10.1016/S0163-6383(84)80022-3

Yeung, H. H., and Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition* 113, 234–243. doi: 10.1016/j.cognition.2009.08.010

Yeung, H. H., and Werker, J. F. (2013). Lip movements affect infant audiovisual speech perception. *Psychol. Sci.* 24, 603–612. doi: 10.1177/0956797612458802

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 March 2014; paper pending published: 06 April 2014; accepted: 22 April 2014; published online: 03 June 2014.

Citation: Carbonell KM and Lotto AJ (2014) Speech is not special. . . again. Front. Psychol. ***5****:427. doi: 10.3389/fpsyg.2014.00427*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 June 2014
doi: 10.3389/fpsyg.2014.00534

Distinct cortical locations for integration of audiovisual speech and the McGurk effect

Laura C. Erickson[1,2†], Brandon A. Zielinski[3,4†], Jennifer E. V. Zielinski[3], Guoying Liu[3,5], Peter E. Turkeltaub[2,6], Amber M. Leaver[1,7] and Josef P. Rauschecker[1,3*]

[1] Department of Neuroscience, Georgetown University Medical Center, Washington, DC, USA
[2] Department of Neurology, Georgetown University Medical Center, Washington, DC, USA
[3] Department of Physiology and Biophysics, Georgetown University Medical Center, Washington, DC, USA
[4] Departments of Pediatrics and Neurology, Division of Child Neurology, University of Utah, Salt Lake City, UT, USA
[5] National Institutes of Health, Bethesda, MD, USA
[6] MedStar National Rehabilitation Hospital, Washington, DC, USA
[7] Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Gregor R. Szycik, Hannover Medical School, Germany
Joana Acha, Basque Centre on Cognition, Brain and Language, Spain

****Correspondence:***
Josef P. Rauschecker, Department of Neuroscience, Georgetown University, Medical Center 3970 Reservoir Road NW, New Research Building WP-19, Washington, DC 20007, USA
e-mail: rauschej@georgetown.edu

† Laura C. Erickson and Brandon A. Zielinski have contributed equally to this work.

Audiovisual (AV) speech integration is often studied using the McGurk effect, where the combination of specific incongruent auditory and visual speech cues produces the perception of a third illusory speech percept. Recently, several studies have implicated the posterior superior temporal sulcus (pSTS) in the McGurk effect; however, the exact roles of the pSTS and other brain areas in "correcting" differing AV sensory inputs remain unclear. Using functional magnetic resonance imaging (fMRI) in ten participants, we aimed to isolate brain areas specifically involved in processing congruent AV speech and the McGurk effect. Speech stimuli were composed of sounds and/or videos of consonant–vowel tokens resulting in four stimulus classes: congruent AV speech (AV_{Cong}), incongruent AV speech resulting in the McGurk effect (AV_{McGurk}), acoustic-only speech (A_O), and visual-only speech (V_O). In group- and single-subject analyses, left pSTS exhibited significantly greater fMRI signal for congruent AV speech (i.e., AV_{Cong} trials) than for both A_O and V_O trials. Right superior temporal gyrus, medial prefrontal cortex, and cerebellum were also identified. For McGurk speech (i.e., AV_{McGurk} trials), two clusters in the left posterior superior temporal gyrus (pSTG), just posterior to Heschl's gyrus or on its border, exhibited greater fMRI signal than both A_O and V_O trials. We propose that while some brain areas, such as left pSTS, may be more critical for the *integration* of AV speech, other areas, such as left pSTG, may generate the "corrected" or merged percept arising from conflicting auditory and visual cues (i.e., as in the McGurk effect). These findings are consistent with the concept that posterior superior temporal areas represent part of a "dorsal auditory stream," which is involved in multisensory integration, sensorimotor control, and optimal state estimation (Rauschecker and Scott, 2009).

Keywords: McGurk effect, superior temporal sulcus, dorsal stream, sensorimotor, cross-modal, multisensory, speech

INTRODUCTION

Two distinct sensory signals are seamlessly integrated during typical speech processing: sounds and facial movements. The integration of acoustic and visual speech cues is frequently studied using the McGurk effect (McGurk and MacDonald, 1976), wherein sounds and facial movements are deliberately mismatched to elicit the perception of an entirely different and illusory consonant–vowel (CV) token. One common example is when the sound "ba" is dubbed onto the visual articulation of "ga," an illusory bimodal "McGurk" percept of "da" results. Yet, the precise neural mechanisms governing integration of congruent audiovisual (AV) speech signals and the subtle perceptual shift of the McGurk effect remain unclear.

Numerous neuroimaging (Sams et al., 1991; Jones and Callan, 2003; Sekiyama et al., 2003; Skipper et al., 2007; Bernstein et al., 2008; Benoit et al., 2010; Wiersinga-Post et al., 2010; Irwin et al., 2011; Nath et al., 2011; Nath and Beauchamp, 2012; Szycik et al., 2012) and behavioral studies (Green et al., 1991; Green and Norrix, 1997; Tiippana et al., 2004, 2011; Nahorna et al., 2012) of the McGurk effect have been published, as well as one transcranial magnetic stimulation (TMS) study (Beauchamp et al., 2010). Substantial emphasis has been placed on the importance of the posterior superior temporal cortex (pST), specifically the left posterior superior temporal sulcus (pSTS), in the McGurk effect (Sekiyama et al., 2003; Bernstein et al., 2008; Beauchamp et al., 2010; Benoit et al., 2010; Irwin et al., 2011; Nath et al., 2011; Nath and Beauchamp, 2012; Szycik et al., 2012). However, other brain regions have also been linked to processing McGurk-type stimuli, including frontal (Skipper et al., 2007; Benoit et al., 2010; Irwin et al., 2011), insular (Skipper et al., 2007; Benoit et al., 2010; Szycik et al., 2012), and parietal areas (Jones and Callan, 2003; Skipper et al., 2007; Benoit et al., 2010; Wiersinga-Post et al., 2010),

as well as other regions (Skipper et al., 2007; Bernstein et al., 2008; Wiersinga-Post et al., 2010; Nath et al., 2011; Szycik et al., 2012). While these experiments examine neural processes related to the McGurk effect, the precise role of each brain region implicated in the McGurk effect, particularly within the pST, is still not completely understood.

The neuroanatomical variability associated with the McGurk effect may be explained by variations in experimental design, as well as differing analytical approaches. Previous studies have probed the McGurk effect using a variety of statistical approaches. Examples include direct contrasts between incongruent McGurk speech versus congruent AV speech (Jones and Callan, 2003; Skipper et al., 2007; Bernstein et al., 2008; Benoit et al., 2010; Irwin et al., 2011; Szycik et al., 2012), or correlations between functional magnetic resonance imaging (fMRI) BOLD activity and McGurk percept reports/susceptibility (Benoit et al., 2010; Wiersinga-Post et al., 2010; Nath et al., 2011; Nath and Beauchamp, 2012). However, these approaches do not isolate regions specifically sensitive to AV signals versus unimodal signals, where interactions of auditory and visual sensory input are likely to occur. This suggests that other methods may be needed to further evaluate the neural correlates of the McGurk effect. Others (Calvert and Thesen, 2004; Beauchamp, 2005b; Laurienti et al., 2005; Stein and Stanford, 2008; Goebel and van Atteveldt, 2009) have discussed several ways to statistically identify neural correlates of multisensory integration, such as assessing the conjunction of auditory and visual signals, and examining differential activation magnitude between AV and unimodal signals (max criterion or super-additive approaches). Beauchamp (2005b) specifically showed that application of different statistical contrasts for AV signals compared to unimodal signals affected activation patterns in the temporal lobe, which is highly relevant when examining the neural correlates of the McGurk effect. Thus, the use of a different statistical approach may help to parse out the cortical processing mechanisms behind the McGurk phenomenon.

In the current study, we attempted to tease apart the distinct neural correlates involved in AV processing of congruent AV speech and McGurk speech. In ten participants using fMRI across the whole brain, we chose the max criterion (Beauchamp, 2005b), which identifies AV-processing regions that respond more strongly to AV stimuli relative to both unimodal auditory and visual stimulation alone. This approach allowed us to focus on brain areas optimized specifically for processing bimodal AV speech, rather than those that respond equally well or indiscriminately to bimodal AV and unimodal stimuli. We suggest that this method allowed for the isolation of AV-processing regions most likely to be involved in processing congruent AV speech or the change in perception accompanying the McGurk effect. This statistical approach has been successfully utilized to isolate AV-processing regions in several language studies (van Atteveldt et al., 2004, 2007; Szycik et al., 2008; Barros-Loscertales et al., 2013) and other types of AV studies (Beauchamp, 2005b; Hein et al., 2007; Watson et al., 2014). Since others have raised the issue of high individual anatomical/functional variability concerning the multisensory portion of the STS (Beauchamp et al., 2010; Nath and Beauchamp, 2012), we confirmed our group results in single-subject analyses, accounting for individual differences in gyral anatomy (Geschwind and Levitsky, 1968) and functional localization within pST. We sought to ensure the location of AV function relative to posterior superior temporal gyrus (pSTG), pSTS, and other landmarks within the pST. Distinguishing between the neural correlates related to AV processing of congruent AV speech and AV processing specific to perceptual ambiguity may help to extend ideas of multisensory functions within current sensorimotor models of language (Skipper et al., 2007; Rauschecker and Scott, 2009; Rauschecker, 2011).

MATERIALS AND METHODS

PARTICIPANTS

Ten volunteers (6 females; mean age = 25.72 years, SD = 3.01) contributed data to this study and were consented in accordance with Georgetown University Institutional Review Board. All participants were right-handed, and primary English speakers. Subjects were recruited through advertisement. Telephone screening ensured that all subjects were in good health with no history of neurological disorders, and reported normal hearing and normal or corrected-to-normal vision. Data from all ten participants were used in statistical analysis.

CONSONANT–VOWEL (CV) TOKEN STIMULI

The following American-English CV tokens were recorded and digitized with sound from six volunteers (3 females and 3 males) articulating the following speech sounds: "ba," "ga," "pa," and "ka," using a Panasonic video-recorder and SGI O2 workstation. Audio and video tracks were edited and recombined using Adobe Premiere. In the videos, only the lower half of each speaker's face was visible, minimizing the influence of gaze and facial processing. Four gain-normalized CV token stimulus types of 2 s duration were created for this experiment: 24 acoustic stimuli with the video track removed (unimodal auditory, A_O), 24 video stimuli with the auditory track removed (unimodal visual, V_O), 24 congruent AV stimuli (AV_{Cong}), and 12 incongruent AV McGurk stimuli (AV_{McGurk}). The relatively large number of different stimuli from six separate speakers for each stimulus type (AV_{Cong}, AV_{McGurk}, A_O, V_O) helped to reduce potential repetition effects. A_O stimuli contained only CV token sounds with no video display of corresponding lower facial movements; only a blank screen was shown. V_O stimuli contained a silent video display of lower facial movements during articulation of a CV token with no corresponding sound presented. AV_{Cong} stimuli contained sound and video from the original CV token recording. For example, auditory "ba" and visual "ba" were recorded from the same speaker during congruent, typical AV speech. AV_{McGurk} stimuli were created from combinations of differing sound and video CV token stimuli to produce two robust McGurk illusions (McGurk and MacDonald, 1976; Green et al., 1991; Green and Norrix, 1997). Twelve different McGurk stimuli were produced to reduce potential repetition effects, where each AV_{McGurk} stimulus was created from the same speaker and presented synchronously. The first set of McGurk stimuli consisted of sound "ba" dubbed onto a video of lips articulating "ga," yielding six stimuli conveying the fused perception "da," one for each recorded speaker. The second set of McGurk stimuli consisted of "pa"

audio dubbed onto a video of lips articulating "ka," producing six stimuli with the fused perception of "ta," one for each recorded speaker.

fMRI EXPERIMENT AND PARADIGM

Scans were acquired using a blocked design in a single fMRI session composed of two runs. AV_{Cong} blocks of trials were presented in the first run, and AV_{McGurk} blocks of trials were presented in the second run. A_O and V_O blocks of trial types were presented in both runs. Three block types were presented in a repeated "A–B–A–C" pattern as follows: AV, V_O, AV, A_O. Each block of trials contained only one type of stimuli, i.e., AV, V_O, or A_O. During each block, seven trials of stimuli (AV, A_O, or V_O) were presented continuously and pseudo-randomly at approximately every 2 s. For each stimulus block, two echo-planar imaging (EPI, or "functional") volumes were collected, and the beginning of each EPI volume was separated by 6.5 s. CV token stimuli were 2 s in length. Thus, in order to create a 13 s stimulus block, actual presentation time for any single stimulus was fractionally less than 2 s. At the beginning of each run, three pre-stimulus "dummy" volumes were collected and removed before statistical analysis to allow for steady-state relaxation. Within each run, 20 blocks were presented, and 40 EPI volumes were acquired, consisting of 20 AV, 10 A_O, and 10 V_O volumes. The total number of EPI volumes collected for both AV_{Cong} and AV_{McGurk} runs included: 20 AV_{Cong}, 20 AV_{McGurk}, 20 A_O, and 20 V_O.

In the MR scanner, binaural auditory stimuli were presented using a custom air-conduction sound system with silicone-cushioned headphones (Resonance Technologies, Van Nuys, CA, USA). The level of auditory stimuli was approximately 75–80 dB SPL, assessed using a B&K Precision Sound Level Meter. Videos (visual stimuli) were presented using a Sharp LCD projector (29.97 fps). Stimuli were projected onto a translucent plexiglass rear-projection screen mounted on the MRI head coil, in which subjects viewed the stimuli via a head coil mirror. All stimuli were presented using a Macintosh G3 personal computer running MacStim (David Darby, Melbourne, VIC, Australia).

In the scanner, the participants' instructions were to attend to the presentation of stimuli, and to covertly count instances of a specific target CV token. This orthogonal task was designed to maintain participant attention and compliance. For example, participants were asked to count the number of "ga" stimuli presented during the AV_{Cong} run. Presence of the illusory McGurk perception for these participants was confirmed by repeating the experiment using the same stimuli as presented during the scan on a computer outside of the MR scanner.

MR IMAGING PARAMETERS

Images were acquired using a 1.5 Tesla Siemens Magnetom Vision whole-body scanner at Georgetown University. Each functional run contained 43 EPI volumes (first 3 pre-stimulus volumes were discarded) that were composed of 25 slices with a slice thickness of 4 mm and a gap of 0.4 mm. We used a repetition time (TR) of 6.5 s, acquisition time (TA) of 3 s, echo time (TE) of 40 ms, and flip angle of 90° with a voxel size of 3.75 mm × 3.75 mm × 4.40 mm. A sparse-sampling design was used to minimize the effect of scanner noise, which is often used in audition studies. EPI volumes were timed to capture the optimal hemodynamic response for each block of trials, allowing the presentation of some stimuli in relative quiet between volumes (Hall et al., 1999). High-resolution MPRAGE scans were acquired using a 256-mm^3 field of view, with a voxel size of 1.00 mm × 1.00 mm × 1.41 mm. Study design, stimuli, experimental paradigm, MR imaging parameters, and data collection were developed, performed, and published as part of previous work (Zielinski, 2002).

fMRI DATA ANALYSIS

All statistical tests were performed in 3D volume-space using BrainVoyager QX (*Brain Innovation*) software. MPRAGE and functional images (EPI volumes) were interpolated into Talairach stereotaxic/standard space (Talairach and Tournoux, 1988). Functional images were preprocessed as follows: (1) motion correction using six parameters, (2) temporal high-pass filter including linear trend removal (3 cycles), (3) spatial Gaussian smoothing (6 mm^3), and (4) co-registration with high-resolution MPRAGE images. During motion correction, images were aligned to the first volume in the run. During spatial normalization, images were aligned across runs. This corrected for any differences in head position both within and across runs.

WHOLE-BRAIN GROUP ANALYSIS

Whole-brain group analysis was conducted using a fixed-effects general linear model (GLM); the fixed-effects analysis method has been successfully used in the current literature (Leaver et al., 2009; Chevillet et al., 2011). GLM predictors were used to measure changes in fMRI signal in single voxels (Friston et al., 1995) and were defined by the timing of blocks of trials for the four types of experimental conditions: AV_{Cong}, AV_{McGurk}, A_O, and V_O. *Post hoc* contrasts compared AV and unimodal conditions (A_O and V_O) within each fMRI run. Group analyses were corrected for multiple voxel-wise comparisons using cluster thresholds determined by the Monte Carlo method as implemented in Brain Voyager, which estimated the probability of false positives (Forman et al., 1995).

To evaluate neural responses to congruent AV speech and McGurk speech across the whole brain, we performed two conjunction ($\cap$) contrasts: (1) $AV_{Cong} > A_O \cap AV_{Cong} > V_O$ and (2) $AV_{McGurk} > A_O \cap AV_{McGurk} > V_O$ (where both statements flanking $\cap$ must be true; **Figure 1**; **Table 1**). This type of multisensory comparison corresponds to the "max criterion" method (Beauchamp, 2005b). It is important to note that since no stimulus-absent condition was tested, no statistical comparisons against "rest-baseline" were conducted. Thus, the fMRI signal changes were estimated by relative differences in beta weights. Significant voxels for these conjunction contrasts exhibited greater fMRI signal for the AV condition than for both unimodal conditions ($p_{corr} < 0.001$ and single-voxel threshold $t > 3.4956$, $p < 0.0005$). Whole-brain analyses using Monte Carlo corrections were conducted within a whole-brain mask defined by only those voxels contained within the averaged brain of the current sample (i.e., an average of the skull-stripped MPRAGEs). Mean beta weights and standard errors for each condition are reported across participants for the left pSTS cluster and left pSTG clusters (**Figure 1**). Beta weights for the two left pSTG clusters were averaged first

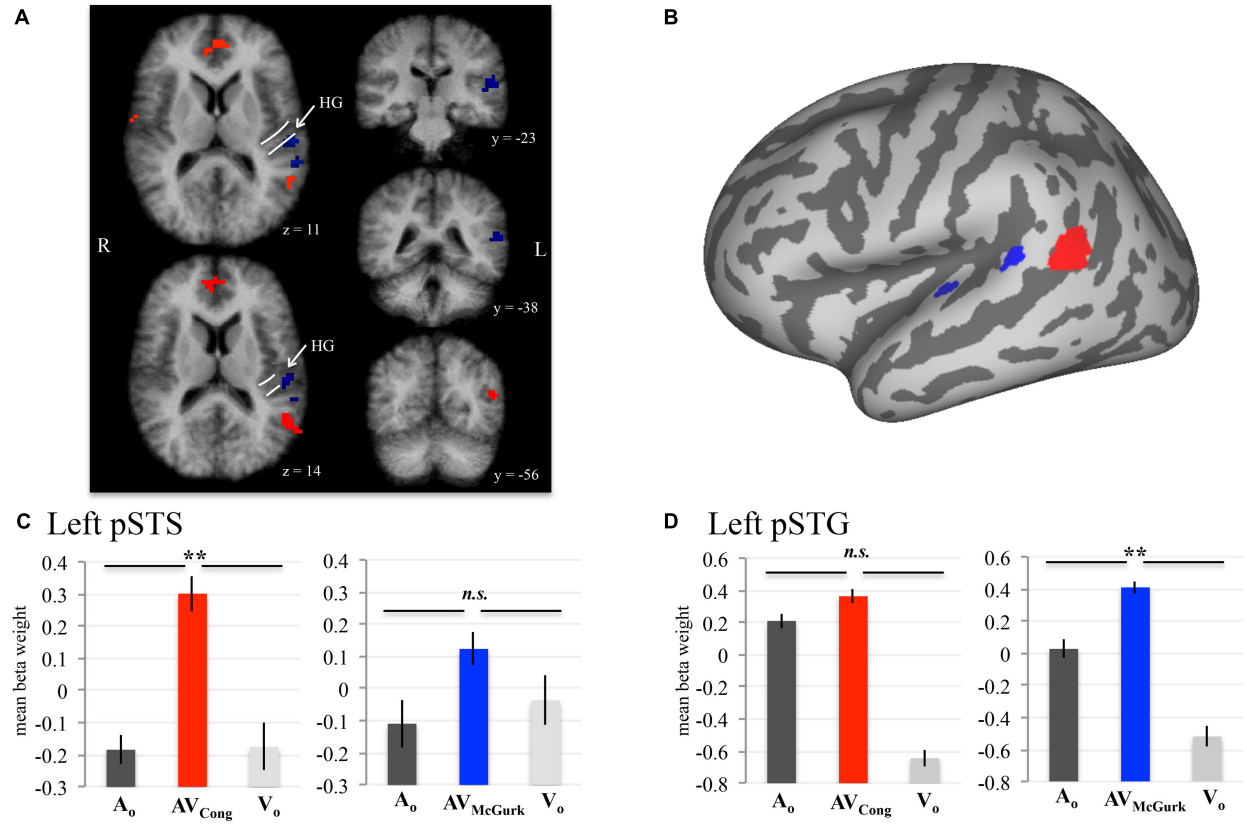


FIGURE 1 | AV speech areas in the left posterior superior temporal cortex for congruent and McGurk speech. Group results ($N = 10$; $p_{corr} < 0.001$) showing voxels with significantly higher fMRI signal for AV speech than both types of unimodal speech (acoustic-only, A_O and visual-only, V_O) are displayed on axial ($z = 11$, 14) and coronal ($y = -23$, –38, –56) 3D volume slices of the averaged brain created from the current sample **(A)**. The anatomic designations were determined in 3D volume space relative to the anatomy on the current sample's averaged brain. The white lines displayed on the axial volume slices approximate the location of HG. Results presented in 3D volume are not interpolated and are presented in radiological convention. The inflated cortical surface template shown in **(B)**, used for display purposes, was not created from the current sample. A conjunction analysis demonstrated that activity in left pSTS (red) was significantly greater in AV_{Cong} trials than in A_O and V_O trials. Two clusters in left pSTG (blue) exhibited a similar pattern for McGurk speech (i.e., $AV_{McGurk} > A_O \cap AV_{McGurk} > V_O$). **(C,D)** Mean fMRI signal for the left pSTS and left pSTG clusters are represented with mean beta weights for AV_{Cong} (red), AV_{McGurk} (blue), A_O (dark gray), and V_O (light gray) blocks of trials. Beta weights for the left pSTG clusters are first averaged across both clusters in each participant. Error bars denote standard error of the mean across participants, and asterisks (**) mark statistically significant effects in the voxel-wise analysis ($p_{corr} < 0.001$). Abbreviations: HG = Heschl's gyrus, pSTS = posterior superior temporal sulcus, pSTG = posterior superior temporal gyrus, n.s. = not significant.

in each participant for every condition, then averaged across participants for the mean beta weight value and standard error. Anatomical location designations of these results were determined based on the anatomy of the averaged brain created from the current sample ($N = 10$) in 3D volume space. These locations were not based on the anatomy of the inflated template cortical surface (**Figure 1B**), which was used only for data presentation and did not reflect the precise anatomy of the current sample.

SINGLE-SUBJECT ANALYSIS IN SUPERIOR TEMPORAL CORTEX

Group findings were confirmed using identical contrasts in single-subject analyses (single-voxel threshold $t > 2.2461$, $p < 0.025$; **Figure 2**), because our sample size may not be optimal for random-effects analysis (Petersson et al., 1999a,b), and fixed-effects analysis does not consider subject variability. To identify single-subject activity that best approximated group findings for either congruent AV speech (on or nearby left pSTS) or McGurk speech (on or nearby left pSTG), we selected voxel(s)/cluster(s) significant for each contrast within the left middle to posterior superior temporal cortex on each participant's brain volume, although other activations (e.g., in temporal cortex) may have been present as well (data not shown). If multiple clusters were chosen for a given subject, then we reported the center of gravity across all clusters together for that participant and mean beta weights were extracted individually from each cluster and averaged for that subject. We validated this selection process by calculating the average Euclidean distance between group and single-subject clusters across participants, using the center of gravity in 3D volume-space.

"MASKED" ANALYSES RESTRICTED TO SENSORY CORTICES

To assess neural responses to congruent AV speech and McGurk speech within auditory and visual cortical regions not detected

Table 1 | Whole-brain group conjunction results ($N = 10$; AV > A_O ∩ AV > V_O) are reported for congruent AV and McGurk speech.

Brain region	Talairach			Volume (mm^3)
	X	*Y*	*Z*	
Congruent AV speech				
Left pSTS	−53	−56	15	621
Right STG	59	−3	5	459
Medial prefrontal cortex	4	46	9	1998
Cerebellum	−3	−49	−21	432
McGurk speech				
Left pSTG	−52	−23	12	810
Left pSTG	−57	−38	12	324

Talairach coordinates represent the center of gravity for each cluster, rounded to the nearest whole number ($p_{corr} < 0.001$).

in whole-brain analysis (**Figure 3**), we created auditory and visual cortex masks from within the averaged brain of the current sample. Auditory cortex was defined by a mask within superior temporal lobe that contained voxels surviving either of two conjunction (∩) contrasts: $AV_{Cong} > V_O \cap A_O > V_O$, or $AV_{McGurk} > V_O \cap A_O > V_O$. The visual cortex mask was created in a similar way using contrasts: $AV_{Cong} > A_O \cap V_O > A_O$ and $AV_{McGurk} > A_O \cap V_O > A_O$. The visual mask included areas within lateral occipital cortex (LOC), and inferior temporal cortex (ITC) containing fusiform gyri. The medial occipital cortex was not included in the mask since A_O trials had slightly higher fMRI signal compared to V_O trials. This does not preclude medial occipital cortex activation in V_O trials; only stimulus-absent trials could confirm this, which were not conducted in this study. To be included in auditory or visual masks, voxels were significant for these contrasts in a whole-brain analysis with a $p_{corr} < 0.001$ determined by single-voxel threshold of $t > 3.9110$, $p < 0.0001$ and displayed with a strict single-voxel threshold of $t > 5.7940$, $p < 1.0 \times 10^{-8}$. AV_{Cong} and AV_{McGurk} effects on masked auditory cortex were defined by two new contrasts: (1) $AV_{Cong} > A_O$, and (2) $AV_{McGurk} > A_O$ ($p_{corr} < 0.01$; single-voxel threshold $t > 1.9630$, $p < 0.05$). AV_{Cong} and AV_{McGurk} effects on masked visual cortex were defined by two new contrasts: (1) $AV_{Cong} > V_O$, and (2) $AV_{McGurk} > V_O$ ($p_{corr} < 0.01$; single-voxel threshold $t > 1.9630$, $p < 0.05$). In other words, significant voxels for these contrasts showed greater fMRI signal for AV trials than for auditory (A_O) trials in masked auditory cortex, or visual (V_O) trials in masked visual cortex. Notably, the contrasts used to define each sensory cortex mask were different from the contrasts used to investigate the bimodal effects in that sensory cortex mask (Kriegeskorte et al., 2009).

DATA PRESENTATION

For visualization purposes, group statistics were exported onto an inflated template cortical surface (Van Essen, 2005), using Caret software (Van Essen et al., 2001) or presented on volume slices of the current sample's averaged brain using BrainVoyager QX (**Figure 1A**). Caret software was used to display foci projections (via "Project Foci to PALS Atlas") onto an inflated template cortical surface for each single-subject result of statistical tests and corresponding centers of gravity (**Figure 2A**). Additionally, single-subject inflated cortical surfaces were constructed using Freesurfer software (Dale et al., 1999; Fischl et al., 1999). Four representative single-subject results (i.e., center of gravity of single-subject analyses, see sub-section *Single-Subject Analysis*) were projected onto their respective individual inflated cortical surfaces in Freesurfer ("mni2tal"; Brett et al., 2002; **Figure 2B**). One subject's data resulted in suboptimal surface reconstruction in some cortical

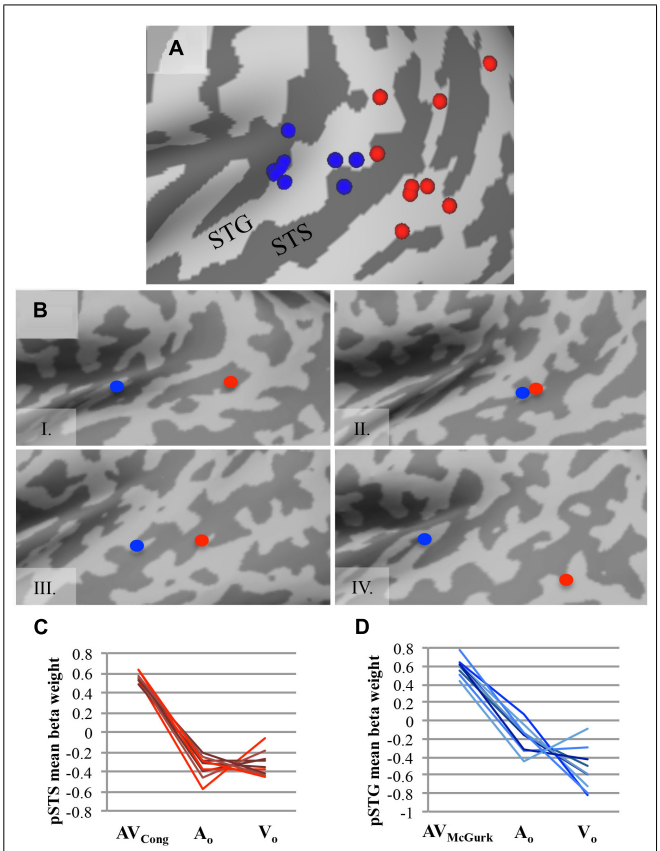


FIGURE 2 | AV speech areas for congruent and McGurk speech in the left posterior superior temporal cortex: consistency across single subjects. (A) The center of gravity for clusters with significantly greater fMRI signal for AV trials than for both types of unimodal trials (AV > A_O ∩ AV > V_O) are plotted for each participant on an inflated template cortical surface (single-voxel threshold $t > 2.2461$, $p < 0.025$). Although the functional location varies, nine out of ten participants exhibited a significant AV effect for AV_{Cong} speech in the left pSTS region (red), and all participants but one exhibited significant AV effects for AV_{McGurk} speech in the left pSTG region (blue). **(B)** Four representative single-subject cortical surface maps are displayed with each participant's foci projected on the inflated surface. Foci represent the center of gravity for clusters with significantly greater fMRI signal for AV speech compared to unimodal speech (AV > A_O ∩ AV > V_O). Red dots represent AV_{Cong} speech, and blue dots represent AV_{McGurk} speech. In most participants, the congruent AV speech clusters were located more posteriorly compared to the McGurk speech clusters. For **(A)** and **(B)**, only the left posterior superior temporal cortex (also including lateral fissure) is displayed. **(C,D)** For each participant, mean fMRI signal from each cluster, estimated by mean beta weight and identified from single-subject activation maps, is depicted for AV and unimodal trials in the left pSTS region **(C)** and the left pSTG region **(D)**; each line represents a single participant. Abbreviations as in **Figure 1**.

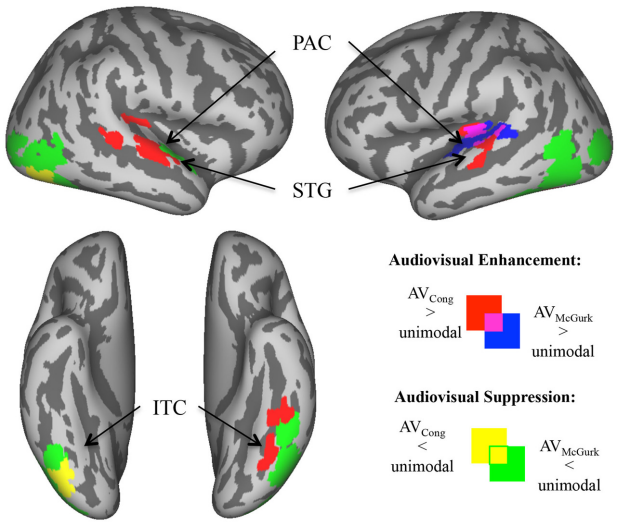


FIGURE 3 | Enhanced and suppressed activity of sensory cortex by congruent AV and McGurk speech. In analyses restricted to auditory and visual cortex (via functionally defined masks), voxels exhibiting significantly greater/enhanced or lesser/suppressed activity in AV blocks of trials as compared to A_O blocks of trials (in auditory cortex) or V_O blocks of trials (in visual cortex) are displayed on inflated template cortical surfaces ($p_{corr} < 0.01$). AV_{Cong} speech (red) compared to unimodal speech had greater fMRI signal in bilateral PAC and mid-STG, and in left ITC including the fusiform gyrus. AV_{McGurk} speech (blue) compared to unimodal speech had greater fMRI signal limited to left PAC and pSTG. By contrast, there were regions of sensory cortex where AV speech had lower fMRI signal compared to unimodal speech. AV_{Cong} speech (yellow) had lower fMRI signal compared to unimodal speech only in right inferior LOC/ITC, whereas during AV_{McGurk} speech (green) stimulation this effect was widely exhibited in bilateral LOC/ITC, and in right ant- to mid-STG. Abbreviations: PAC = primary auditory cortex, ant-STG = anterior superior temporal gyrus, mid-STG = middle superior temporal gyrus, pSTG = posterior superior temporal gyrus, ITC = inferior temporal cortex, LOC = lateral occipital cortex.

areas, but tissue segmentation was accurate in the superior temporal cortex; thus it did not affect the assessment of individual anatomy within this region.

RESULTS

BRAIN AREAS INVOLVED IN AV PROCESSING OF CONGRUENT SPEECH

Brain areas associated with processing congruent AV speech were identified from the comparison of the fMRI signal on blocks of trials containing AV recordings of congruent CV stimuli (AV_{Cong}) to blocks of trials including only unimodal CV stimuli (A_O and V_O) across the whole brain. The left pSTS exhibited activation where fMRI signal for AV_{Cong} trials was significantly greater than both A_O and V_O trials (red; **Figure 1**; $p_{corr} < 0.001$ for conjunction contrast: $AV_{Cong} > A_O \cap AV_{Cong} > V_O$). Three other brain areas were found: right STG, medial prefrontal cortex, and cerebellum (**Table 1**). In summary, regions identified here, including the left pSTS, have increased response to congruent AV versus unimodal sensory input compared to other areas in the whole brain.

BRAIN AREAS INVOLVED IN AV PROCESSING OF MCGURK SPEECH

Brain areas involved in processing McGurk speech, composed of incongruent acoustic and visual signals, were identified from the comparison of fMRI signal on blocks of trials containing incongruent McGurk-type AV recordings of CV stimuli (AV_{McGurk}) to blocks of trials containing only unimodal CV stimuli (A_O and V_O) across the whole brain (blue; **Figure 1**). Two adjacent clusters were identified in left pSTG, located just posterior to Heschl's gyrus. It is possible that one of these McGurk clusters may be on the border of Heschl's gyrus (–52, –23, 12). The anatomical designation of pSTG was based on the anatomy of the current sample's averaged brain in 3D volume space. These left pSTG clusters exhibited activation where fMRI signal for AV_{McGurk} trials was significantly greater than both A_O and V_O trials ($p_{corr} < 0.001$ for conjunction contrast: $AV_{McGurk} > A_O \cap AV_{McGurk} > V_O$). Increased response to McGurk speech compared to unimodal sensory signals was only identified in regions of the left pSTG.

SINGLE-SUBJECT CONFIRMATION OF pST REGIONS INVOLVED IN PROCESSING CONGRUENT AV AND MCGURK SPEECH

To confirm the effects found in the group analysis, single-subject analyses were conducted to locate brain areas more responsive to AV_{Cong} or AV_{McGurk} trials compared to unimodal speech, A_O and V_O, using the same statistical contrasts described above. Activation within the left pSTS region was identified for congruent AV speech in nine out of ten participants (**Figure 2**; single-voxel threshold $t > 2.2461$, $p < 0.025$), where the fMRI signal for AV_{Cong} trials was greater than both unimodal trials (A_O and V_O). While the exact location of congruent AV speech clusters identified in the left pSTS region varied among participants, in general, clusters reported here were positioned on the left pSTS or neighboring regions, nearby or overlapping with the group left pSTS finding. These clusters were typically posterior to the individual clusters identified for McGurk speech. However, some participants also showed activation for congruent AV speech in regions similar to the regions identified during McGurk speech (**Figure 2B**). One subject did not show activation to congruent AV speech in left pSTS; however, this subject did show an effect for McGurk speech in left pSTG. The individual locations of congruent AV speech areas differed from the group cluster in the left pSTS by an average of 10.91 ± SD 5.52 mm. The locations of these clusters were carefully determined relative to individual anatomy through evaluations in both volume and in individual surface reconstructions of pST (**Figure 2**).

Recruitment of the left pSTG region was confirmed in processing McGurk speech in single-subject analyses in nine out of ten participants (single-voxel threshold $t > 2.2461$, $p < 0.025$; **Figure 2**), where the fMRI signal for AV_{McGurk} trials was greater than both unimodal trials (A_O and V_O), i.e., using the same conjunction contrast as in the whole-brain group analysis. Individual locations of activation in the pSTG region differed among participants, but in general were positioned on the pSTG or surrounding cortex (e.g., adjacent STS) and were near to or overlapped with the group left pSTG findings. While one participant did not exhibit this effect in left pSTG, this subject did demonstrate the effect in left pSTS for congruent AV speech. The single-subject centers of gravity of fMRI signal compared to the McGurk speech group foci in left pSTG varied by 11.91 ± SD 3.47 mm, averaged for both left pSTG group clusters in each individual, further indicating that there may be individual differences in functional location.

Single-subject activations typically overlapped with one or both of the two McGurk group clusters, suggesting that each cluster may likely represent a focal point of activation within the larger area of left pSTG, perhaps extending into Heschl's gyrus, rather than two areas with distinct functions.

ENHANCED ACTIVITY IN SENSORY CORTEX BY AV SPEECH

Areas of enhanced activity were localized within masked auditory and visual cortex, where AV blocks of trials exhibited greater fMRI signal compared to unimodal A_O blocks of trials in auditory cortex ($AV > A_O$) or V_O blocks of trials in visual cortex ($AV > V_O$). In sensory cortex, congruent AV speech (red; **Figure 3**) had greater fMRI signal compared to unimodal speech bilaterally in primary auditory cortex (PAC) extending into mid-superior temporal gyri (mid-STG), and in left ITC including the fusiform gyrus ($p_{corr} < 0.01$). We consider PAC to be located in medial Heschl's gyrus (Morosan et al., 2001). In contrast, McGurk speech (blue; **Figure 3**) had greater fMRI signal compared to unimodal speech solely in left PAC spreading into pSTG ($p_{corr} < 0.01$). Overlap of these effects for both congruent AV speech and McGurk speech were localized within the left PAC and pSTG, similar to some single-subject results. In general, these results show that different regions within sensory cortex exhibit preference to congruent AV speech and McGurk speech, complementing results reported above from whole-brain group analyses.

SUPPRESSED ACTIVITY IN SENSORY CORTEX BY AV SPEECH

Within masked auditory and visual sensory cortex, some regions exhibited significantly lower fMRI signal for AV speech blocks of trials compared to unimodal A_O blocks of trials in auditory cortex ($AV < A_O$) or V_O blocks of trials in visual cortex ($AV < V_O$). Activity in these areas of sensory cortex revealed a higher fMRI signal to unimodal speech compared to AV speech. Congruent AV speech (yellow; **Figure 3**) demonstrated lower fMRI signal compared to unimodal trials only in right inferior LOC/ITC ($p_{corr} < 0.01$). This effect was not detected in auditory cortex. In contrast, McGurk speech (green; **Figure 3**) broadly exhibited lower fMRI signal compared to unimodal trials, including right anterior to middle superior temporal gyrus (ant-STG), and bilateral LOC/ITC ($p_{corr} < 0.01$).

DISCUSSION

Whole-brain group analyses ($N = 10$) that were confirmed in single-subject analyses suggested that distinct posterior superior temporal regions are involved in processing congruent AV and McGurk speech when compared to unimodal speech (acoustic-only and visual-only). Left pSTS was recruited when processing congruent bimodal AV speech, suggesting that this region may be speech-sensitive and critical when sensory signals converge to be compared. In contrast, left pSTG was recruited when processing McGurk speech, suggesting that left pSTG may be necessary when discrepant auditory and visual cues interact. We interpret these findings as suggesting that two similar neural processes take place in separate left pST regions: (1) comparison and integration of sensory cues in the left pSTS and (2) creation of the "corrected" or merged percept in the left pSTG arising from conflicting auditory and visual cues. In other words, a new merged percept is generated in pSTG, resulting from the incorporation of conflicting auditory and visual speech cues. It is possible that alternate interpretations may explain these findings. Future studies will need to more closely examine the precise role of these regions (left pSTG vs. left pSTS) related to general AV-integrative processes. In general, these findings help to support and refine current sensorimotor models of speech processing, especially with regard to multisensory interactions in posterior superior temporal cortex (Skipper et al., 2007; Rauschecker and Scott, 2009; Rauschecker, 2011).

AV INTEGRATION IN THE LEFT pSTS

The left pSTS was recruited during congruent AV speech, which suggests a general AV-processing function that could support integration of auditory and visual speech signals. The idea that the pSTS is important for multisensory integration (Beauchamp, 2005a; Beauchamp et al., 2008), particularly AV integration of language (Calvert et al., 2000; Beauchamp et al., 2004a; van Atteveldt et al., 2004; Stein and Stanford, 2008; Nath and Beauchamp, 2011) and other stimuli (Beauchamp et al., 2004b; Noesselt et al., 2007; Hein and Knight, 2008; Man et al., 2012; Powers et al., 2012; Watson et al., 2014), is not new. In a recent example, Man et al. (2012) demonstrated similar neural activity patterns in the left pSTS for non-speech visual-only representation and acoustic-only representation of the same object. Supporting our findings, the left pSTS has been consistently recruited in AV language studies using the max criterion for AV integration (conjunction of $AV > A_O$ and $AV > V_O$; Beauchamp, 2005b) of congruent AV stimuli including various stimulus types, such as sentences in native and non-native language (Barros-Loscertales et al., 2013), words (Szycik et al., 2008), and visual letters paired with speech sounds (van Atteveldt et al., 2004, 2007). Similarly, the left pSTS showed increased activity to congruent AV story stimuli compared to the sum of activity for acoustic-only and visual-only stimulation (Calvert et al., 2000); others have also reported supra-additive AV speech effects in STS (Wright et al., 2003). Evidence that the STS is involved in processing many kinds of sensory input (Hein and Knight, 2008), such as biological motion (Grossman and Blake, 2002) and socially relevant sensory cues (Allison et al., 2000; Lahnakoski et al., 2012), further suggests a general sensory integration function. Our findings and others (Beauchamp et al., 2004a; Man et al., 2012) support the possibility that the pSTS could be responsible for a more general, non-exclusive AV function that compares and integrates AV sensory cues.

Previous studies implicate the left pSTS in the McGurk effect (Sekiyama et al., 2003; Beauchamp et al., 2010; Benoit et al., 2010; Nath et al., 2011; Nath and Beauchamp, 2012). However, these studies do not imply an exclusive role of the left pSTS in the McGurk percept change *per se*. For example, activity in the STS does not always have a strong response to McGurk syllables in some children who have high McGurk percept likelihood (Nath et al., 2011) or a preference to McGurk stimuli over other incongruent AV stimuli in adults (Nath and Beauchamp, 2012). In Japanese speakers, the left pSTS was recruited more during noisy McGurk trials compared to noise-free McGurk trials

(Sekiyama et al., 2003), which may reflect an increased demand for AV integration rather than specificity for the McGurk perceptual shift. Further, while inhibitory TMS of the left pSTS significantly decreased the prevalence of reported McGurk percepts, some other AV-influenced percepts were still produced, e.g., "between 'ba' and 'da'," "b-da," or new percept "ha," albeit at a much lower incidence (Beauchamp et al., 2010). This suggests that part of the mechanism responsible for changing or "correcting" the auditory percept based on AV signals is still intact after inactivation of left pSTS. Finally, it is worth noting that left pSTS can be recruited by incongruent (not McGurk stimuli) more than by congruent AV stimuli (Zielinski, 2002; Bernstein et al., 2008; Hocking and Price, 2008; Szycik et al., 2009), perhaps suggesting the left pSTS is involved in situations of incongruence beyond the McGurk effect. Considering our findings in the context of previous work, we suggest that left pSTS may be necessary for the McGurk effect by virtue of its role in general AV processing; however, we suggest the possibility that the resulting change in perception famous to the McGurk effect may occur elsewhere.

CREATION OF "CORRECTED" PERCEPTS IN THE LEFT pSTG

Our data show that two clusters in the left pSTG (just posterior to Heschl's gyrus based on the current sample's averaged brain) were recruited by McGurk speech. One interpretation of our findings is that the left pSTG may have a role in generating new "corrected" percepts underlying the McGurk effect. In other words, pSTG creates a new merged percept by incorporating input from conflicting auditory and visual cues reflective of both streams of information. Previous research, including some McGurk studies, supports this interpretation. One study using pattern analysis in the pSTG and posterior auditory regions was able to decode differences in percept, either "aba" or "ada," when presented with identical AV stimuli, suggesting that the pSTG is sensitive to perception and not just acoustics (Kilian-Hutten et al., 2011; cf. Chevillet et al., 2013). Despite limited previous evidence, other studies have indicated auditory areas including the pSTG in the McGurk effect (Skipper et al., 2007; Benoit et al., 2010; Szycik et al., 2012), especially where assessments focused on the neural correlates and/or fMRI time courses associated with the change in McGurk speech percept, or the visual modulation present in the McGurk effect. Supporting our findings, Szycik et al. (2012) identified left pSTG activation during McGurk trials when participants reported the McGurk percept and when comparing participants who perceived the McGurk effect to those who did not. Although these pSTG areas are discussed as left "pSTS," we speculate that it is possible these areas may be on the left pSTG with Talairach foci reported close to the center of gravity of the pSTG clusters identified in our study (our congruent AV pSTS cluster was further posterior). Benoit et al. (2010) showed an adaptation effect for McGurk stimuli in bilateral middle to posterior STG extending into pSTS when the sound was held constant while the visual cue changed, reflecting the auditory perceptual change due to visual influence. Finally, Skipper et al. (2007) provided evidence for percept changes in auditory and somatosensory areas, where early versus late fMRI time courses for McGurk stimuli displayed different neural activation patterns that correlated more to congruent AV "pa" or "ta," respectively. Building on these previous findings, we propose that, during the McGurk effect, the left pSTG may have a more specific function in generating auditory percepts incorporating the influence of multiple sensory modalities.

AV ENHANCEMENT AND SUPPRESSION OF ACTIVITY IN SENSORY CORTICES AND OTHER REGIONS

Differential AV responses for congruent AV and McGurk speech are further supported when examining enhancement (increases) and suppression (decreases) of activity in auditory and visual sensory cortex by AV speech compared to acoustic-only or visual-only speech. During congruent AV speech, AV enhancement occurred throughout auditory and visual areas, whereas AV suppression was limited to right LOC. LOC has been previously linked to face/object processing (Grossman and Blake, 2002) and biological motion processing (Vaina et al., 2001). The seeming suppression of the LOC in the right hemisphere in the current study could be related to the left-lateralization of speech/language processes. Similarly, in the main analysis, the right STG had increased activity when comparing congruent AV speech to both acoustic-only and visual-only speech. These results may be due to imagery (Driver, 1996; Kraemer et al., 2005; Zatorre and Halpern, 2005), attention effects (Grady et al., 1997; Pekkola et al., 2006; Tiippana et al., 2011), and/or increased overall input during AV speech compared to only acoustic or visual speech (Hocking and Price, 2008). In contrast, McGurk speech enhancement was only identified in the left pSTG and PAC, and overall there was more AV suppression of auditory and visual sensory cortex. It is possible that the left pSTG and PAC were the only sensory sites benefiting from AV input during McGurk speech, or it could be that these areas process incongruent AV input differently than the rest of sensory cortex. In either case, comparing the relatively widespread enhancement and limited suppression of sensory cortical activity during congruent AV speech to the more circumscribed enhancement of left posterior auditory areas and extensive suppression of sensory cortex during McGurk speech further underscores a potential specialized role of the pSTG in generating auditory percepts reflective of the conflicting AV input present during the McGurk effect.

Although we have focused primarily on the posterior superior temporal cortex, other brain regions are involved in analyzing and integrating AV speech as well. This is exemplified during congruent AV speech, where other regions recruited include medial prefrontal cortex and cerebellum. Medial prefrontal cortex activation has been demonstrated in speech comprehension (Obleser et al., 2007) and recent meta-analytic evidence (Zald et al., 2014) showed consistent coactivation of the adjacent medial and lateral orbitofrontal cortex and the left pST region. The left pSTS and medial prefrontal cortex may process information specific to emotion category (anger, *etc.*), independent of whether the input is received from facial movements, body movements, or the voice (Peelen et al., 2010). Likewise, cerebellum may be involved in speech processing (Sekiyama et al., 2003; Skipper et al., 2005; Ackermann, 2008; Wiersinga-Post et al., 2010), as well as processing music (Leaver et al., 2009). The cerebellum has also been implicated in visual processes related to biological motion, e.g., where biological motion was depicted by visual point-light displays of various human movements (Grossman et al., 2000).

Future work is needed to address the interplay and functional relationships between different brain regions during typical AV speech perception. It is important to note that AV interactions not only lead to enhancement of activity; they can also accelerate the detection of visual change in speech, as measured with magnetoencephalography (Möttönen et al., 2002).

ALTERNATE INTERPRETATIONS AND LIMITATIONS

Alternate interpretations of these findings are possible. For example, AV information may be integrated differently depending on the composition of the AV signal. The processing differences related to integration of McGurk speech could solely result from incongruent auditory and visual sensory inputs and not necessarily from a perceptual change. Similarly, McGurk speech may simply contribute more sensory information than congruent AV speech, where processing of incongruent McGurk speech could have an increased 'load' (see Hocking and Price, 2008). However, these interpretations are unlikely because others have found the STS to be activated by McGurk stimuli (Sekiyama et al., 2003; Beauchamp et al., 2010; Benoit et al., 2010; Nath et al., 2011; Nath and Beauchamp, 2012), and other incongruent AV stimuli (Zielinski, 2002; Bernstein et al., 2008; Hocking and Price, 2008; Szycik et al., 2009), suggesting that the STS can process multiple types of AV information including incongruent AV sensory cues. Thus, it is possible that the left pSTG may be involved in a different neural process, such as changing auditory percepts based on the integration of differing auditory and visual cues that are present during McGurk speech. Future experiments are needed to examine bimodal vs. unimodal comparisons with incongruent AV speech stimuli that do not elicit a McGurk or other illusory percepts.

It is also possible that the group findings for McGurk speech in the pSTG extend onto Heschl's gyrus, because there was variability in the location of the McGurk speech clusters in single-subject analyses, and one of the group McGurk clusters may be on the border of Heschl's gyrus. The McGurk clusters may overlap with regions equivalent to lateral belt or parabelt areas in non-human primates (Rauschecker et al., 1995; Kaas and Hackett, 2000; Hackett, 2011); however, because these regions are not yet defined with sufficient precision in the human brain (but see Chevillet et al., 2011), the level of auditory processing recruited during McGurk speech is unclear. Thus, if earlier auditory areas including regions of Heschl's gyrus are recruited during processing of McGurk speech, this would suggest that the "corrected" McGurk percept may be created at an earlier processing stage. Future experiments can further test for perceptual change processes in different regions of the pSTG extending to primary or core auditory areas.

We should note that this experiment also had other limitations. First, while the reported effects in left pSTS and pSTG were identified in whole-brain group analyses and confirmed in single-subject analyses, these results were derived from a relatively small sample ($N = 10$), indicating a slightly lower power than with the standard minimum of $N = 12$ (Desmond and Glover, 2002). Furthermore, the McGurk percept was confirmed in our participants outside of the scanner, in order to limit participant motion, which means the presence of the McGurk effect during the scan is largely inferred. In general, future studies with a larger number of participants are needed to confirm the possibility of differential multisensory effects related to congruent AV speech and the perceptual change associated with the McGurk effect in the pST.

CONCLUSION: THE MCGURK EFFECT AND THE AUDITORY DORSAL STREAM

Our main findings reveal that the left pSTS may have a more general function in AV processing and the left pSTG may be more involved in processing AV perceptual change. These results have the potential to inform current ideas regarding multisensory function and organization of the pST, particularly in consideration of sensorimotor models of speech processing (Skipper et al., 2007; Rauschecker and Scott, 2009; Rauschecker, 2011). To focus on one model, Rauschecker and Scott (2009) expanded the current dual-stream auditory theory (Rauschecker and Tian, 2000) and proposed that dorsal-stream regions, including the pST, are involved in sensorimotor interactions and multisensory processes. They suggest that these functions may be related to speech and other "doable" sounds, which may facilitate error reduction and "disambiguation of phonological information." Our findings support this model and further suggest that differential AV interactions within the pST may contribute to these sensorimotor transformations and comparisons. The idea that the McGurk effect may be composed of two neural processes of AV integration and "percept correction," complements a similar behavioral model, in which the McGurk effect is a two-stage process of "binding and fusion" (Nahorna et al., 2012). In conclusion, we suggest the possibility that the left pSTG and pSTS may have separate functions, wherein the left pSTG may be specially involved in "correcting" incongruent percepts and the left pSTS may function to integrate congruent AV signals.

AUTHOR CONTRIBUTIONS

All authors meet all four criteria required of authorship. Brandon A. Zielinski and Josef P. Rauschecker conceived of and designed the study; Brandon A. Zielinski, Jennifer E. V. Zielinski, and Guoying Liu conducted data acquisition; Laura C. Erickson, Amber M. Leaver, and Brandon A. Zielinski conducted data analysis; Laura C. Erickson, Brandon A. Zielinski, Peter E. Turkeltaub, Amber M. Leaver, and Josef P. Rauschecker conducted data interpretation; Laura C. Erickson, Amber M. Leaver, Brandon A. Zielinski, and Josef P. Rauschecker wrote the manuscript; all authors critically reviewed the manuscript.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Nos. DGE-0903443 and DGE-1444316 to Laura C. Erickson and National Science Foundation Grant Nos. BCS-0519127 and OISE-0730255 to Josef P. Rauschecker. This work was also supported by National Institutes of Health Grant Nos. R01 EY018923 and R01 NS052494 to Josef P. Rauschecker; T32 NS041231 also funded Laura C. Erickson; NRSA Individual Predoctoral Fellowship 1F31MH012598 and CHRCDA K12HD001410 to Brandon A. Zielinski, as well as the Primary Children's Medical Center Foundation (Early Career Development Award) to Brandon A. Zielinski.

REFERENCES

Ackermann, H. (2008). Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives. *Trends Neurosci.* 31, 265–272. doi: 10.1016/j.tins.2008.02.011

Allison, T., Puce, A., and McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends Cogn. Sci.* 4, 267–278. doi: 10.1016/S1364-6613(00)01501-1

Barros-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Avila Rivera, C., et al. (2013). Neural correlates of audiovisual speech processing in a second language. *Brain Lang.* 126, 253–262. doi: 10.1016/j.bandl.2013.05.009

Beauchamp, M. S. (2005a). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr. Opin. Neurobiol.* 15, 145–153. doi: 10.1016/j.conb.2005.03.011

Beauchamp, M. S. (2005b). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics* 3, 93–113. doi: 10.1385/NI:3:2:093

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., and Martin, A. (2004a). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.* 7, 1190–1192. doi: 10.1038/nn1333

Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004b). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41, 809–823. doi: 10.1016/S0896-6273(04)00070-4

Beauchamp, M. S., Nath, A. R., and Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *J. Neurosci.* 30, 2414–2417. doi: 10.1523/JNEUROSCI.4865-09.2010

Beauchamp, M. S., Yasar, N. E., Frye, R. E., and Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *Neuroimage* 41, 1011–1020. doi: 10.1016/j.neuroimage.2008.03.015

Benoit, M. M., Raij, T., Lin, F. H., Jääskeläinen, I. P., and Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Hum. Brain Mapp.* 31, 526–538. doi: 10.1002/hbm.20884

Bernstein, L. E., Lu, Z. L., and Jiang, J. (2008). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Res.* 1242, 172–184. doi: 10.1016/j.brainres.2008.04.018

Brett, M., Johnsrude, I. S., and Owen, A. M. (2002). The problem of functional localization in the human brain. *Nat. Rev. Neurosci.* 3, 243–249. doi: 10.1038/nrn756

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10, 649–657. doi: 10.1016/S0960-9822(00)00513-3.

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

Chevillet, M., Riesenhuber, M., and Rauschecker, J. P. (2011). Functional correlates of the anterolateral processing hierarchy in human auditory cortex. *J. Neurosci.* 31, 9345–9352. doi: 10.1523/JNEUROSCI.1448-11.2011

Chevillet, M. A., Jiang, X., Rauschecker, J. P., and Riesenhuber, M. (2013). Automatic phoneme category selectivity in the dorsal auditory stream. *J. Neurosci.* 33, 5208–5215. doi: 10.1523/JNEUROSCI.1870-12.2013

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9, 179–194. doi: 10.1006/nimg.1998.0395

Desmond, J. E., and Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J. Neurosci. Methods* 118, 115–128. doi: 10.1016/S0165-0270(02)00121-8

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* 381, 66–68. doi: 10.1038/381066a0

Fischl, B., Sereno, M. I., and Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9, 195–207. doi: 10.1006/nimg.1998.0396

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.* 33, 636–647. doi: 10.1002/mrm.1910330508

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S., et al. (1995). Analysis of fMRI time-series revisited. *Neuroimage* 2, 45–53. doi: 10.1006/nimg.1995.1007

Geschwind, N., and Levitsky, W. (1968). Human brain: left-right asymmetries in temporal speech region. *Science* 161, 186–187. doi: 10.1126/science.161.3837.186

Goebel, R., and van Atteveldt, N. (2009). Multisensory functional magnetic resonance imaging: a future perspective. *Exp. Brain Res.* 198, 153–164. doi: 10.1007/s00221-009-1881-7

Grady, C. L., Van Meter, J. W., Maisog, J. M., Pietrini, P., Krasuski, J., and Rauschecker, J. P. (1997). Attention-related modulation of activity in primary and secondary auditory cortex. *Neuroreport* 8, 2511–2516. doi: 10.1097/00001756-199707280-00019

Green, K. P., Kuhl, P. K., Meltzoff, A. N., and Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Percept. Psychophys.* 50, 524–536. doi: 10.3758/BF03207536

Green, K. P., and Norrix, L. W. (1997). Acoustic cues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions. *J. Speech Lang. Hear. Res.* 40, 646–665.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *J. Cogn. Neurosci.* 12, 711–720. doi: 10.1162/089892900562417

Grossman, E. D., and Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron* 35, 1167–1175. doi: 10.1016/S0896-6273(02)00897-8

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hear. Res.* 271, 133–146. doi: 10.1016/j.heares.2010.01.011

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223. doi: 10.1002/(SICI)1097-0193(1999)7:3<213::AID-HBM5>3.0.CO;2-N

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J. Neurosci.* 27, 7881–7887. doi: 10.1523/JNEUROSCI.1740-07.2007

Hein, G., and Knight, R. T. (2008). Superior temporal sulcus – It's my area: or is it? *J. Cogn. Neurosci.* 20, 2125–2136. doi: 10.1162/jocn.2008.20148

Hocking, J., and Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cereb. Cortex* 18, 2439–2449. doi: 10.1093/cercor/bhn007

Irwin, J. R., Frost, S. J., Mencl, W. E., Chen, H., and Fowler, C. A. (2011). Functional activation for imitation of seen and heard speech. *J. Neuroling.* 24, 611–618. doi: 10.1016/j.jneuroling.2011.05.001

Jones, J. A., and Callan, D. E. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport* 14, 1129–1133. doi: 10.1097/01.wnr.0000074343.81633.2a

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799. doi: 10.1073/pnas.97.22.11793

Kilian-Hutten, N., Valente, G., Vroomen, J., and Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *J. Neurosci.* 31, 1715–1720. doi: 10.1523/JNEUROSCI.4572-10.2011

Kraemer, D. J., Macrae, C. N., Green, A. E., and Kelley, W. M. (2005). Musical imagery: sound of silence activates auditory cortex. *Nature* 434, 158. doi: 10.1038/434158a

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., et al. (2012). Naturalistic FMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Front. Hum. Neurosci.* 6:233. doi: 10.3389/fnhum.2012.00233

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., and Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp. Brain Res.* 166, 289–297. doi: 10.1007/s00221-005-2370-2

Leaver, A. M., Van Lare, J., Zielinski, B., Halpern, A. R., and Rauschecker, J. P. (2009). Brain activation during anticipation of sound sequences. *J. Neurosci.* 29, 2477–2485. doi: 10.1523/JNEUROSCI.4921-08.2009

Man, K., Kaplan, J. T., Damasio, A., and Meyer, K. (2012). Sight and sound converge to form modality-invariant representations in temporoparietal cortex. *J. Neurosci.* 32, 16629–16636. doi: 10.1523/JNEUROSCI.2342-12.2012

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., and Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage* 13, 684–701. doi: 10.1006/nimg.2000.0715

Möttönen, R., Krause, C. M., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cogn. Brain Res.* 13, 417–425. doi: 10.1016/S0926-6410(02)00053-8

Nahorna, O., Berthommier, F., and Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* 132, 1061–1077. doi: 10.1121/1.4728187

Nath, A. R., and Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J. Neurosci.* 31, 1704–1714. doi: 10.1523/JNEUROSCI.4853-10.2011

Nath, A. R., and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787. doi: 10.1016/j.neuroimage.2011.07.024

Nath, A. R., Fava, E. E., and Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *J. Neurosci.* 31, 13963–13971. doi: 10.1523/JNEUROSCI.2605-11.2011

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441. doi: 10.1523/JNEUROSCI.2252-07.2007

Obleser, J., Wise, R. J., Dresner, M. A., and Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *J. Neurosci.* 27, 2283–2289. doi: 10.1523/JNEUROSCI.4663-06.2007

Peelen, M. V., Atkinson, A. P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *J. Neurosci.* 30, 10127–10134. doi: 10.1523/JNEUROSCI.2161-10.2010

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., and Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. *Hum. Brain Mapp.* 27, 471–477. doi: 10.1002/hbm.20190

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. (1999a). Statistical limitations in functional neuroimaging. *I.* Non-inferential methods and statistical models. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1239–1260. doi: 10.1098/rstb.1999.0477

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. (1999b). Statistical limitations in functional neuroimaging. *II.* Signal detection and statistical inference. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1261–1281. doi: 10.1098/rstb.1999.0478

Powers, A. R. III, Hevey, M. A., and Wallace, M. T. (2012). Neural correlates of multisensory perceptual learning. *J. Neurosci.* 32, 6263–6274. doi: 10.1523/JNEUROSCI.6138-11.2012

Rauschecker, J. P. (2011). An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hear. Res.* 271, 16–25. doi: 10.1016/j.heares.2010.09.001

Rauschecker, J. P., and Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* 12, 718–724. doi: 10.1038/nn.2331

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111–114. doi: 10.1126/science.7701330

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-f.

Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neurosci. Res.* 47, 277–287. doi: 10.1016/S0168-0102(03)00214-1.

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89. doi: 10.1016/j.neuroimage.2004.11.006

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 255–266. doi: 10.1038/nrn2331

Szycik, G. R., Jansma, H., and Münte, T. F. (2009). Audiovisual integration during speech comprehension: an fMRI study comparing ROI-based and whole brain analyses. *Hum. Brain Mapp.* 30, 1990–1999. doi: 10.1002/hbm.20640

Szycik, G. R., Stadler, J., Tempelmann, C., and Münte, T. F. (2012). Examining the McGurk illusion using high-field 7 Tesla functional MRI. *Front. Hum. Neurosci.* 6:95. doi: 10.3389/fnhum.2012.00095

Szycik, G. R., Tausche, P., and Münte, T. F. (2008). A novel approach to study audiovisual integration in speech perception: localizer fMRI and sparse sampling. *Brain Res.* 1220, 142–149. doi: 10.1016/j.brainres.2007.08.027

Talairach, J., and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging.* Stuttgart; New York: Georg Thieme.

Tiippana, K., Andersen, T. S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

Tiippana, K., Puharinen, H., Möttönen, R., and Sams, M. (2011). Sound location can influence audiovisual speech perception when spatial attention is manipulated. *Seeing Perceiving* 24, 67–90. doi: 10.1163/187847511X-557308

Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., and Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11656–11661. doi: 10.1073/pnas.191374198

van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron* 43, 271–282. doi: 10.1016/j.neuron.2004.06.025

van Atteveldt, N. M., Formisano, E., Blomert, L., and Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cereb. Cortex* 17, 962–974. doi: 10.1093/cercor/bhl007

Van Essen, D. C. (2005). A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex. *Neuroimage* 28, 635–662. doi: 10.1016/j.neuroimage.2005.06.058

Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., and Anderson, C. H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. *J. Am. Med. Inform. Assoc.* 8, 443–459. doi: 10.1136/jamia.2001.0080443

Watson, R., Latinus, M., Charest, I., Crabbe, F., and Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex* 50, 125–136. doi: 10.1016/j.cortex.2013.07.011

Wiersinga-Post, E., Tomaskovic, S., Slabu, L., Renken, R., De Smit, F., and Duifhuis, H. (2010). Decreased BOLD responses in audiovisual processing. *Neuroreport* 21, 1146–1151. doi: 10.1097/WNR.0b013e328340cc47

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., and McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex* 13, 1034–1043. doi: 10.1093/cercor/13.10.1034

Zald, D. H., McHugo, M., Ray, K. L., Glahn, D. C., Eickhoff, S. B., and Laird, A. R. (2014). Meta-analytic connectivity modeling reveals differential functional connectivity of the medial and lateral

orbitofrontal cortex. *Cereb. Cortex* 24, 232–248. doi: 10.1093/cercor/bhs308

Zatorre, R. J., and Halpern, A. R. (2005). Mental concerts: musical imagery and auditory cortex. *Neuron* 47, 9–12. doi: 10.1016/j.neuron.2005.06.013

Zielinski, B. A. (2002). Auditory-Visual Interactions in the Perception of Species-Specific Communication Sounds in the Human: Towards a Comprehensive Model of Elementary Sound Processing in Primates. Ph.D. thesis, Georgetown University, Washington, DC, published through UMI, Ann Arbor, MI.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 February 2014; accepted: 14 May 2014; published online: 02 June 2014.
Citation: Erickson LC, Zielinski BA, Zielinski JEV, Liu G, Turkeltaub PE, Leaver AM and Rauschecker JP (2014) Distinct cortical locations for integration of audiovisual speech and the McGurk effect. Front. Psychol. ***5****:534. doi: 10.3389/fpsyg.2014.00534*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 November 2014
doi: 10.3389/fpsyg.2014.01340

A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception

Attigodu C. Ganesh[1]*, Frédéric Berthommier[1], Coriandre Vilain[1], Marc Sato[2] and Jean-Luc Schwartz[1]*

[1] CNRS, Grenoble Images Parole Signal Automatique-Lab, Speech and Cognition Department, UMR 5216, Grenoble University, Grenoble, France
[2] CNRS, Laboratoire Parole et Langage, Brain and Language Research Institute, UMR 7309, Aix-Marseille University, Aix-en-Provence, France

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Mireille Besson, Centre National de la Recherche Scientifique – Institut de Neurosciences Cognitives de la Mediterranée, France
Victoria Knowland, City University, UK

***Correspondence:**
Attigodu C. Ganesh and Jean-Luc Schwartz, CNRS, Grenoble Images Parole Signal Automatique-Lab, Speech and Cognition Department, UMR 5216, Grenoble University, 11 rue des Mathématiques, Grenoble Campus BP46, F-38402, Saint Martin d'Heres cedex, Grenoble, France
e-mail: ganesh.attigodu@gipsa-lab.grenoble-inp.fr; jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

Audiovisual (AV) speech integration of auditory and visual streams generally ends up in a fusion into a single percept. One classical example is the McGurk effect in which incongruent auditory and visual speech signals may lead to a fused percept different from either visual or auditory inputs. In a previous set of experiments, we showed that if a McGurk stimulus is preceded by an incongruent AV context (composed of incongruent auditory and visual speech materials) the amount of McGurk fusion is largely decreased. We interpreted this result in the framework of a two-stage "binding and fusion" model of AV speech perception, with an early AV binding stage controlling the fusion/decision process and likely to produce "unbinding" with less fusion if the context is incoherent. In order to provide further electrophysiological evidence for this binding/unbinding stage, early auditory evoked N1/P2 responses were here compared during auditory, congruent and incongruent AV speech perception, according to either prior coherent or incoherent AV contexts. Following the coherent context, in line with previous electroencephalographic/magnetoencephalographic studies, visual information in the congruent AV condition was found to modify auditory evoked potentials, with a latency decrease of P2 responses compared to the auditory condition. Importantly, both P2 amplitude and latency in the congruent AV condition increased from the coherent to the incoherent context. Although potential contamination by visual responses from the visual cortex cannot be discarded, our results might provide a possible neurophysiological correlate of early binding/unbinding process applied on AV interactions.

Keywords: audiovisual binding, speech perception, multisensory interactions, EEG

INTRODUCTION

Speech perception requires adequate hearing and listening skills, but it is well known that visual information from the face and particularly from lip movements may intervene in the speech decoding process. The first classical evidence for audiovisual (AV) integration in speech perception in normal-hearing subjects concerns the role of lip reading during speech comprehension, with a gain in the AV modality in respect to the audio-only modality particularly in adverse listening conditions (e.g., Sumby and Pollack, 1954; Erber, 1971; Benoît et al., 1994; Grant and Seitz, 2000; Bernstein et al., 2004b). Another classical behavioral example for AV integration is provided by the McGurk effect (McGurk and MacDonald, 1976), in which a conflicting visual input modifies the perception of an auditory input (e.g., visual /ga/ added on auditory /ba/ leading to the percept of /da/). This led researchers to propose a number of possible architectures for AV integration, according to which auditory and visual information converge toward a single percept in the human brain (Massaro, 1987; Summerfield, 1987; Schwartz et al., 1998).

A number of studies have then searched for potential neurophysiological and neuroanatomical correlates of AV integration in speech perception. At the neurophysiological level, recent electroencephalographic (EEG) and magnetoencephalographic (MEG) studies focused on the influence of the visual input on the auditory event-related potentials (ERPs), notably on auditory N1 (negative peak, occurring typically 100 ms after the sound onset) and P2 (positive peak, occurring typically 200 ms after the sound onset) responses considered to be associated with the processing of the physical and featural attributes of the auditory speech stimulus prior to its categorization (Näätänen and Winkler, 1999). In the last 10 years, various studies consistently displayed an amplitude reduction of N1/P2 auditory responses together with a decrease in their onset latency. These studies typically involved consonant–vowel syllables uttered in isolation, with a natural advance of the visual input (associated with the phonation preparation) on the sound. Their results suggest that the visual input modulates and speeds up the neural processing of auditory ERPs as soon as 100 ms after the sound onset and that AV integration partly occurs at an early processing stage in the cortical auditory speech processing hierarchy (Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Pilling, 2009; Vroomen and Stekelenburg, 2010; Baart et al., 2014; Knowland et al., 2014; Treille et al., 2014a,b). The interpretation has generally called upon "predictive mechanisms" (van Wassenhove et al., 2005), according to which the visual input, arriving ahead of sound, would enable to predict part of its content and hence modulate the auditory ERP in amplitude

and latency. The visual modulation seems to obey different rules respectively for N1 and P2. For N1, it would just depend on the advance of the image over the sound, even for incongruent auditory and visual inputs, and even for non-speech stimuli; while the P2 modulation would be speech specific and crucially depend on the phonetic content of the auditory and visual inputs (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010).

While the AV integration process has long been considered as automatic (e.g., Massaro, 1987; Soto-Faraco et al., 2004), a number of recent papers have provided evidence that it could actually be under the control of attentional processes (e.g., Tiippana et al., 2004; Alsius et al., 2005, 2007; Colin et al., 2005; Navarra et al., 2005; Mozolic et al., 2008; Buchan and Munhall, 2012). Furthermore, previous results on the "AV speech detection advantage" (Grant and Seitz, 2000; Kim and Davis, 2004) and its consequences for AV perception (Schwartz et al., 2004) suggest a mechanism by which early visual processing would reduce spectral and temporal uncertainty in the auditory flow. This mechanism, thought to operate prior to AV fusion, would detect whether the visual and acoustic information are bound to the same articulatory event and should be processed together. This view, reinforced by electrophysiological data on early AV speech interactions, suggest that AV interactions could intervene at various stages in the speech decoding process (Bernstein et al., 2004a).

In a similar vein, Berthommier (2004) proposed that AV fusion could rely on a two-stage process, beginning by binding together the appropriate pieces of auditory and visual information, followed by integration *per se* (**Figure 1**). The binding stage would occur early in the AV speech processing chain enabling the listener to extract and group together the adequate cues in the auditory and visual streams, exploiting coherence in the dynamics of the sound and sight of the speech input. In **Figure 1**, the binding stage is displayed by the output of the "coherence" box assessing the likelihood that the audio and video inputs are indeed associated to the same speech event. The output of the binding stage would provide the input to a second processing stage where categorization (and possibly detection in the AV speech detection paradigm) would occur. Integration would hence occur only at this second stage, and conditioned both by general attentional processes but also by the result of the binding stage. If AV coherence is low, binding is unlikely and integration should be weaker.

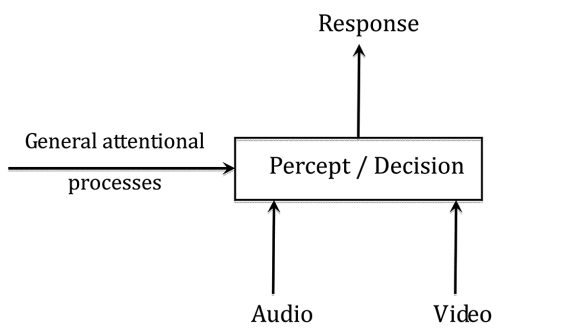


FIGURE 1 | A two-stage "binding and fusion" model of audiovisual speech perception.

To attempt to demonstrate the existence of this "binding" process, Berthommier and colleagues defined an experimental paradigm possibly leading to "unbinding" (Nahorna et al., 2012). In this paradigm (see **Figure 2**), incongruent "McGurk" (A/ba/ + V/ga/) or congruent "ba" (A/ba/ + V/ba/) targets were preceded by congruent or incongruent AV contexts (to distinguish incongruence in context and in targets, we use the terms "coherent" and "incoherent" for context in the following). The expectation was that the incoherent context should induce the subjects to decrease their confidence that the auditory and visual streams were related to a coherent source. This should decrease the role of the visual input on phonetic decision and hence result in a decrease of the McGurk effect. This is what they called "unbinding." The experimental results supported this hypothesis. Indeed, compared to the coherent contexts, various kinds of incoherent contexts, such as acoustic syllables dubbed on video sentences, or phonetic or temporal modifications of the acoustic content of a regular sequence of AV syllables, produced significant amounts of reduction in the McGurk effect. In line with the two-stage model of AV fusion (see **Figure 1**), these results suggest that fusion can be conditioned by prior contexts on AV coherence. They also appear compatible with the above-cited behavioral data on AV detection suggesting that the coherence of the auditory and visual inputs is computed early enough to enhance auditory processing, resulting in the AV speech detection advantage.

The present study aimed at determining a possible neurophysiological marker of the AV binding/unbinding process in the cortical auditory speech hierarchy. Capitalizing on the results obtained by Nahorna et al. (2013), the experiment was adapted from previous EEG experiments on AV speech perception, adding either a coherent or an incoherent AV context before auditory, congruent AV and incongruent AV speech stimuli. The assumption is that with coherent context we should replicate the results of previous EEG studies on auditory N1/P2 responses (decrease in amplitude and latency in the AV vs. A condition). However, an incoherent context should lead to unbinding, as in Nahorna et al. (2013), with the consequence that the visual influence on the auditory stimulus should decrease. Hence the N1/P2 latency and amplitude in the AV condition should increase (reaching a value close to their value in the A condition) in the incoherent context compared with the coherent context.

MATERIALS AND METHODS

PARTICIPANTS

Nineteen healthy volunteers (17 women and 2 men, mean age = 30 years, SD = 13.1 years) participated in the experiment. All participants were French native speakers (although no standard tests were used to measure first or, possibly, second language proficiency), right-handed, without any reported history of hearing disorders and with normal or corrected-to-normal vision. Written consent was obtained from each participant and all procedures were approved by the Grenoble Ethics Board (CERNI). The participants were paid for participating in the experiment.

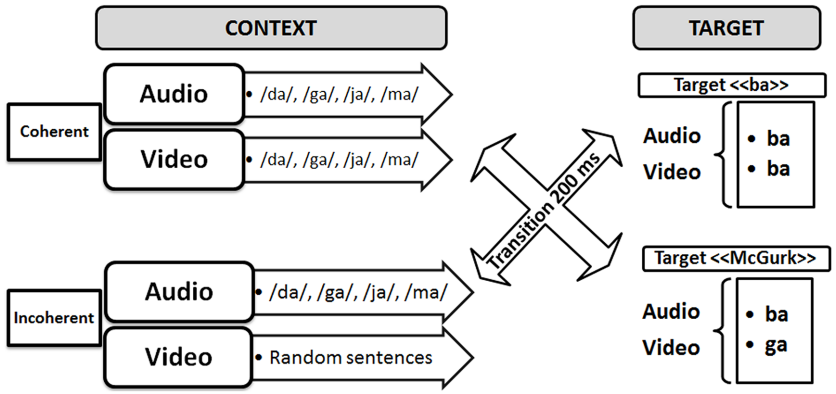


FIGURE 2 | Experimental paradigm for displaying possible "binding/unbinding" mechanisms where context would modulate the McGurk effect.

STIMULI

The audio–video stimuli were similar to those of the previous experiments by Nahorna et al. (2012, 2013) that is with an initial part called "context" followed by a second part called "target." The target was either a pure audio stimulus ("pa" or "ta" dubbed with a fixed image for the same duration), or a congruent AV stimulus ("pa" or "ta") or an incongruent "McGurk" stimulus (audio "pa" dubbed on a video "ka"). The AV context was either coherent or incoherent (**Figure 3**). Coherent contexts consisted of regular sequences of coherent AV syllables randomly selected within the following syllables ("va," "fa," "za," "sa," "ra," "la," "ja," "cha," "ma," "na"). These syllables were selected within the set of possible /Ca/ syllables in French, where C is a consonant not contained in the /p t k b d g/ set, so that target syllables /pa ta ka/ or their perceptually close voiced counterparts /ba da ga/, cannot appear in the context. In the incoherent context material, the auditory content was the same, but the visual content was replaced by excerpts of video sentences, produced in a free way by the same speaker, and matched in duration. The context and target, both of fixed duration (respectively 2 and 1.08 s), were separated by a 1 s period of silence and fixed black image.

All stimuli were prepared from two sets of AV material, a "syllable" material and a "sentence" material, produced by a French male speaker, with lips painted in blue to allow precise video analysis of lip movements (Lallouache, 1990). Videos were edited in Adobe Premier Pro into a 720/576 pixel movie with a digitization rate of 25 frames/s (1 frame = 40 ms). Stereo soundtracks were digitized in Adobe Audition at 44.1 kHz with 16-bit resolution.

The duration of each trial was 5280 ms, in which the context AV movie, lasting 2000 ms, was followed by silence for 1000 ms, then by the target with a duration of 1080 ms. The response time was 1200 ms. To ensure continuity between the end of the context stimulus and silence and also between silence and the onset of the target stimulus, a 120-ms transition stimulus was included

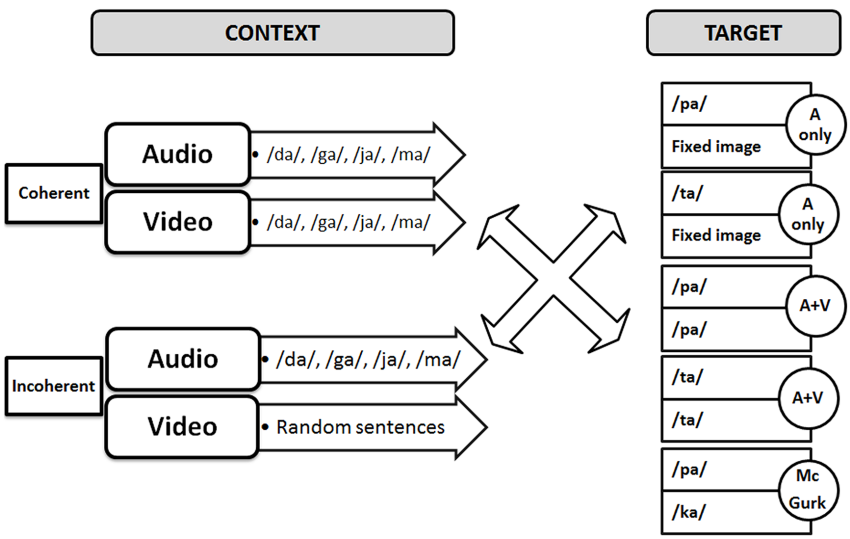


FIGURE 3 | Audiovisual material used in the experiment.

by image fusion (see **Figure 4**). Video fade-in and fade-out were also included in the first and last three frames, respectively. In the auditory only conditions, the auditory targets were presented with a static face of the speaker. The difference between the visual and auditory onsets for /pa/ and /ta/ were respectively 287 and 206 ms.

PROCEDURE

The subject's task was to categorize the stimuli as "pa" or "ta," by pressing the appropriate key (two-alternative forced-choice identification task). Stimulus presentation was coordinated with the Presentation software (Neurobehavioral Systems). In order to avoid possible interference between speech identification and motor response induced by key pressing participants were told to produce their responses a short delay after the stimulus end when a question mark symbol appeared on the screen (typically 320 ms after the end of the stimulus). There were six conditions, with three targets (audio-only, A vs. AV congruent, AVC vs. AV incongruent, AVI) and two contexts (coherent vs. incoherent), and altogether 100 repetitions per condition (with 50 "pa" and 50 "ta" in the audio-only or AV congruent targets, and 100 McGurk stimuli). This provided altogether 600 occurrences, presented in a random order inside five experimental blocks. Altogether, the experiment lasted more than 1 h, including subject preparation, explanations and pauses between blocks. This unfortunately removed the possibility to add a specific visual-only condition, since it would have added two targets – visual congruent and visual incongruent – and hence almost doubled the experiment duration. We will discuss in various parts of the paper what the consequences of this specific choice could be in the processing and interpretation of EEG data.

The experiment was carried out in a soundproof booth with the sound presented through a loudspeaker at a comfortable and fixed level for all subjects. The video stream was displayed on a screen at a rate of 25 images per second, the subject being positioned at about 50 cm from the screen. Participants were instructed to categorize each target syllable by pressing on one key corresponding to /pa/ or /ta/ on a computer keyboard (with a counterbalanced order between subjects) with their left hand.

EEG PARAMETERS

Electroencephalography data were continuously recorded from 64 scalp electrodes (Electro-Cap International, Inc., according to the international 10–20 system) using the Biosemi ActiveTwo AD-box EEG system operating at a 256 Hz sampling rate. Two additional electrodes served as reference [common mode sense (CMS) active electrode] and ground [driven right leg (DRL) passive electrode]. One other external reference electrode was put at the top of the nose. Electro-oculogram measures of the horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes at the outer canthus of each eye as well as above and below the right eye. Before the experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

ANALYSES

All EEG data were processed using the EEGLAB toolbox (Delorme and Makeig, 2004) implemented in Matlab (Mathworks, Natick, MA, USA). EEG data were first re-referenced off-line to the nose recording and band-pass filtered using a two-way least-squares FIR filtering (2–20 Hz). Data were then segmented into epochs of 600 ms including a 100 ms pre stimulus baseline, from –100 to 0 ms to the acoustic target syllable onset, individually determined for each stimulus from prior acoustical analyses). Epochs with an amplitude change exceeding $\pm 100\ \mu V$ at any channel (including HEOG and VEOG channels) were rejected (<5%).

As previously noted, because of time limitations a visual-alone condition was not incorporated in the study, while it is generally included in EEG studies on AV perception. However, to attempt to rule out the possibility that visual responses from the occipital areas could blur and contaminate auditory evoked responses in fronto-central electrodes, we performed various topography analyses using EEGLAB to define the spatial distributions and dynamics of the activity on the scalp surface. Fp1, Fz, F2, P10, P9, and Iz electrodes were not included in this analysis because of noisy electrodes or dysfunction of electrodes for at least one participant. We studied the spatial distribution in two steps. Firstly, we plotted the scalp maps for all six conditions (context × modality) to confirm that the maximal N1/P2 auditory evoked potentials were indeed localized around fronto-central sites on the scalp.

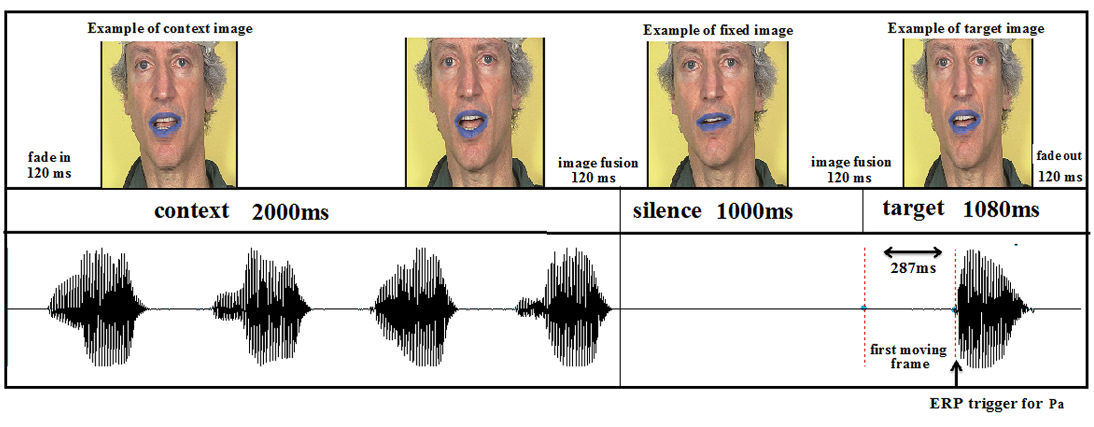


FIGURE 4 | Experimental sequence.

The aim of the second step was to evaluate the presence and amount of possible contamination in the auditory fronto-central electrodes by the visual responses in corresponding cortical areas dedicated to the processing of visual information. To do so, we calculated scalp maps between conditions in the N1-P2 time period.

Since the first part of the topographic analysis confirmed that maximal N1/P2 auditory evoked potentials indeed occurred over fronto-central sites on the scalp (see **Figure 5**; see also Scherg and Von Cramon, 1986; Näätänen and Picton, 1987), and in line with previous EEG studies on AV speech perception and auditory evoked potentials (e.g., van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2009; Vroomen and Stekelenburg, 2010; Treille et al., 2014a,b), an ERP analysis was then conducted on six representative left, middle, and right fronto-central electrodes (F3, Fz, F4, C3, Cz, C4) in which AV speech integration has been previously shown to occur (note that Fz was replaced by the average of F1 and F2 responses for two participants because of a dysfunction of electrodes). For each participant, the peak latencies of auditory N1 and P2 evoked responses were first manually determined on the EEG waveform averaged over all six electrodes for each context and modality. Two temporal windows were then defined on these peaks ±30 ms in order to individually calculate N1 and P2 amplitude and latency for all modalities, context and electrodes. Peak detection was done automatically.

For P2 amplitude and latency it has to be noticed that the N1-to-P2 latency could reach small values as low as 75 ms, with double P2 peaks for many subjects. This is not unclassical: double

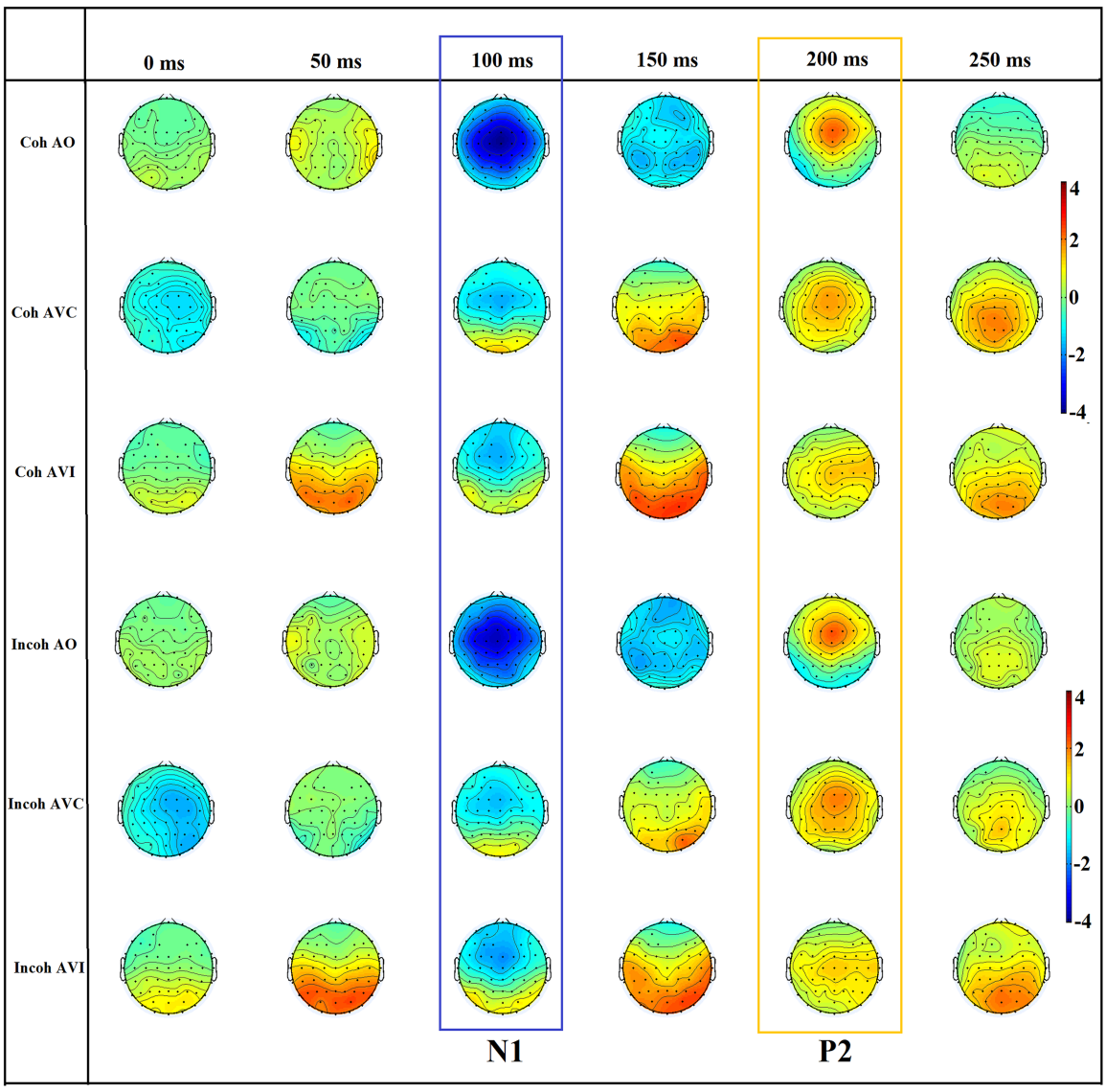


FIGURE 5 | The scalp topography of N1 and P2 for the six conditions (Coh AO, Coh AVC, Coh AVI, Incoh AO, Incoh AVC, Incoh AVI) in time steps of 50 ms. The range of the voltage maps is from –4 to 4 μv.

peaks in the P2 time period have actually been found in a number of studies in both adults, children, elderly and also in impaired populations (e.g., Ponton et al., 1996; Hyde, 1997; Ceponiene et al., 2008; Bertoli et al., 2011). Since the classical range for P2 is 150–250 ms and since the first P2 peak was close to this range, the analysis was focused on the first P2 peak for further analyses.

Notice that we also tested another baseline earlier on in the silence portion between context and target that is from –500 to –400 ms to the acoustic target syllable onset, and we checked that this did not change the results presented later, in any crucial way, either in whole graphs or statistical analysis.

Repeated-measure analyses of variances (ANOVAs) were performed on N1 and P2 amplitude and latency with context (coherent vs. incoherent) and modality (A vs. AVC vs. AVI) as within-subjects variables. Partial eta squared values were systematically provided to estimate effect sizes. *Post hoc* analyses with Bonferroni correction were done when appropriate, and are reported at the $p < 0.05$ level.

Concerning behavioral data, the proportion of responses coherent with the auditory input was individually determined for each participant, each syllable, and each modality. A repeated-measure ANOVA was performed on this proportion with context (coherent vs. incoherent) and modality (A vs. AVC vs. AVI) as within-subjects variables. *Post hoc* analyses with Bonferroni correction were done when appropriate, and are reported at the $p < 0.05$ level.

RESULTS

BEHAVIORAL ANALYSIS

On **Figure 6** we display the behavioral scores, presented as percentage of responses coherent with the auditory input. The scores were close to 100% in the A and AV conditions. They were lower in the AVI conditions, since the visual input changes the percept and produces some McGurk effect. The main effect of modality of presentation was significant [$F(2,36) = 6.14$, $p < 0.0.05$], with more correct responses in A and AVC than in AVI modalities (as shown by *post hoc* analyses; on average, A: 98.2%, AV: 98.3%, and AVI: 77.7%). There was no significant effect of context or interaction. Contrary to our previous studies (Nahorna et al., 2012, 2013), the amount of McGurk effect is hence very small and independent on context. This is likely due to the specific procedure associated with EEG experiments in which the number of different stimuli is quite low (only five different target stimuli altogether) with highly predictable targets.

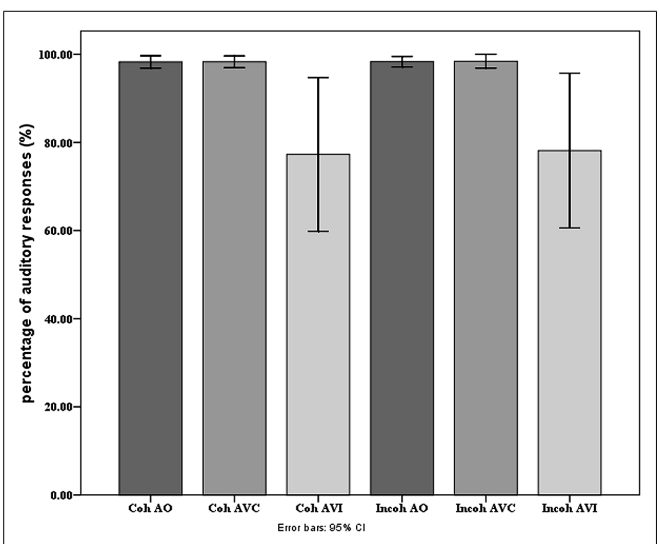


FIGURE 6 | Mean percentage of responses coherent with the auditory input in each modality and context presentation in the behavioral experiment. Error bars represent standard errors of the mean.

EEG ANALYSES

N1 amplitude and latency (see Figures 7 and 8A,B)

In the following analysis, N1 amplitudes were reported in absolute values, hence reduced amplitude means a reduction in absolute value and an increase in real (negative) values. The repeated-measures ANOVA on N1 amplitude displayed no significant effect of context, but a significant effect of modality [$F(2,36) = 13.29$, $p < 0.001$, $\eta_p^2 = 0.42$], with a reduced N1 amplitude observed for the AVC and AVI modalities as compared to the A modality (**Figure 8A**). The *post hoc* analysis shows that the amplitudes in both AVC (–2.00 μV) and AVI (–1.64 μV) were indeed smaller compared to A (–3.62 μV) irrespective of context. Interaction between context and modality was not significant.

The repeated-measures ANOVA on N1 latency displayed no significant effect of context (**Figure 8B**). The modality effect was close to significance [$F(2,36) = 3.20$, $p = 0.07$, $\eta_p^2 = 0.15$], with a shorter latency in the AVI (109 ms) compared to the A (115 ms) and AVC (115 ms) conditions. Interaction between context and modality was not significant.

In brief the results about N1 amplitude are similar to the previously mentioned EEG studies on AV speech perception, with a visually induced amplitude reduction for both congruent (AVC) and incongruent (AVI) stimuli irrespective of context. Regarding N1 latency, the difference between auditory and AV modalities is smaller than in few previous EEG studies, and consequently not significant.

P2 amplitude and latency (see Figures 7 and 8C,D)

There was no significant effect of context or modality in P2 amplitude, but the interaction between context and modality was significant [$F(2,36) = 3.51$, $p < 0.05$, $\eta_p^2 = 0.16$], which is in line with our hypothesis (**Figure 8C**). To further examine the interaction effect between context and modality in P2 amplitude, pairwise comparisons were done using Bonferroni corrections to test the effect of context separately for each modality. The *post hoc* analysis within modality provided a significant difference between Coherent and Incoherent AVC conditions ($p = 0.01$), showing that Coherent AVC (1.15 μV) has smaller amplitude compared to Incoherent AVC (2.03 μV). Context provided no other significant differences either in the AVI or in the A modality.

Concerning P2 latency (**Figure 8D**), there was a significant effect of context [$F(1,18) = 5.63$, $p < 0.05$, $\eta_p^2 = 0.23$], the latency in the Coherent context (176 ms) being smaller than in the Incoherent context (185 ms). There was also a significant effect of modality [$F(2,36) = 23.35$, $p < 0.001$, $\eta_p^2 = 0.56$], P2 occurring

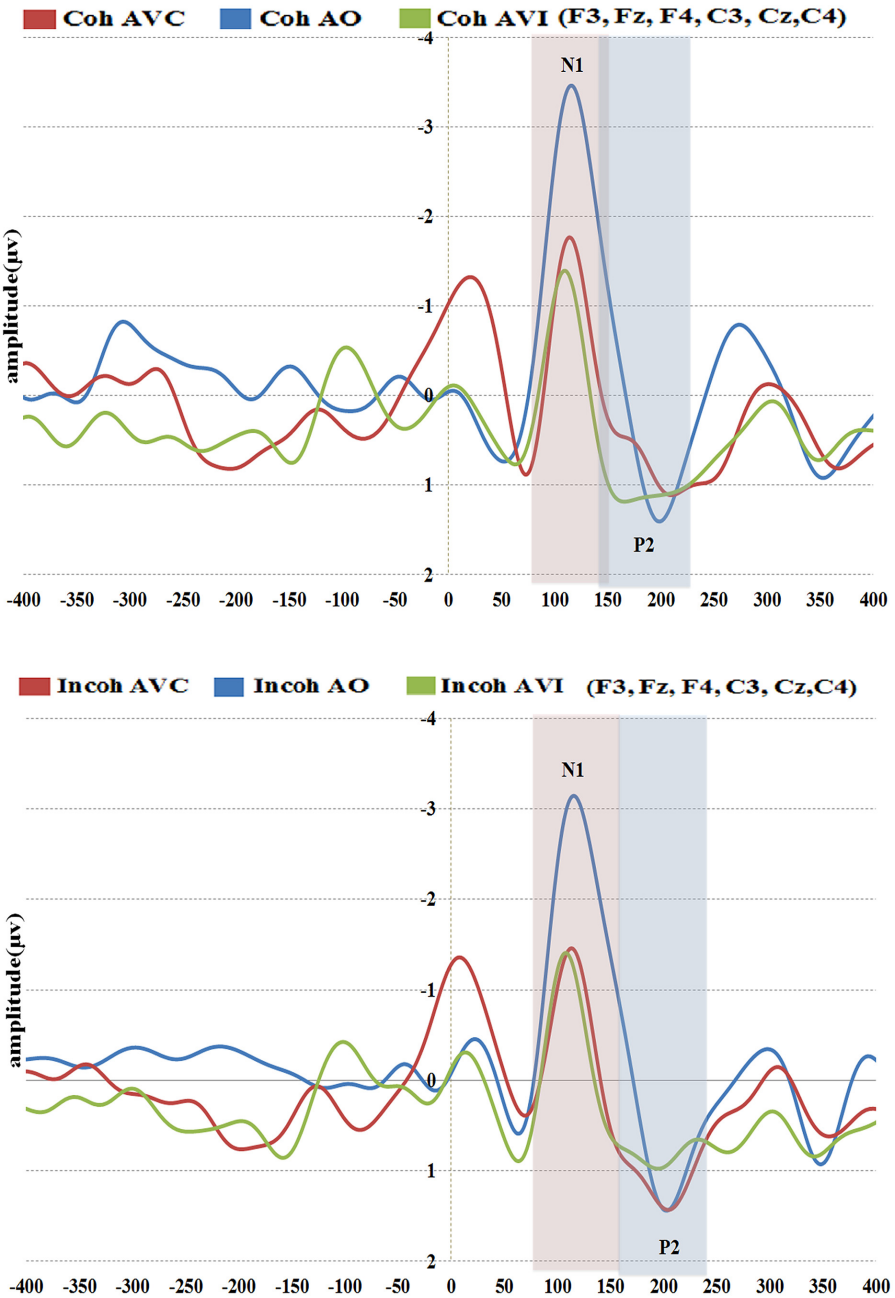


FIGURE 7 | Grand-average of auditory evoked potentials for the six electrodes (frontal and central) for coherent (top) and incoherent (bottom) context and in the three conditions (AO, AVC, and AVI).

earlier in the AVC (178 ms) and AVI (167 ms) modalities compared to AO (196 ms). As in the case of P2 amplitude, there was a significant interaction effect between context and modality [$F(2,36) = 8.07$, $p < 0.005$, $\eta_p^2 = 0.31$]. The *post hoc* analysis provided a significant difference between Coherent and Incoherent AVC conditions ($p = 0.002$), showing that P2 in the Coherent AVC condition occurred earlier (165 ms) than in the Incoherent AVC condition (190 ms). Context provided no other significant differences either in the AVI or in the A modality.

Therefore, contrary to the data for N1, we observed significant effects of context for P2. These effects concern both amplitude and latency. They are focused on the AVC condition with rather large values (25 ms increase in latency and 0.88 μV increase in amplitude from Coherent to Incoherent context in the AVC condition). They result in removing the latency difference between AVC and A, in line with our expectations. However, there appears to be no effect of context in the AVI condition, neither for amplitude nor for latency.

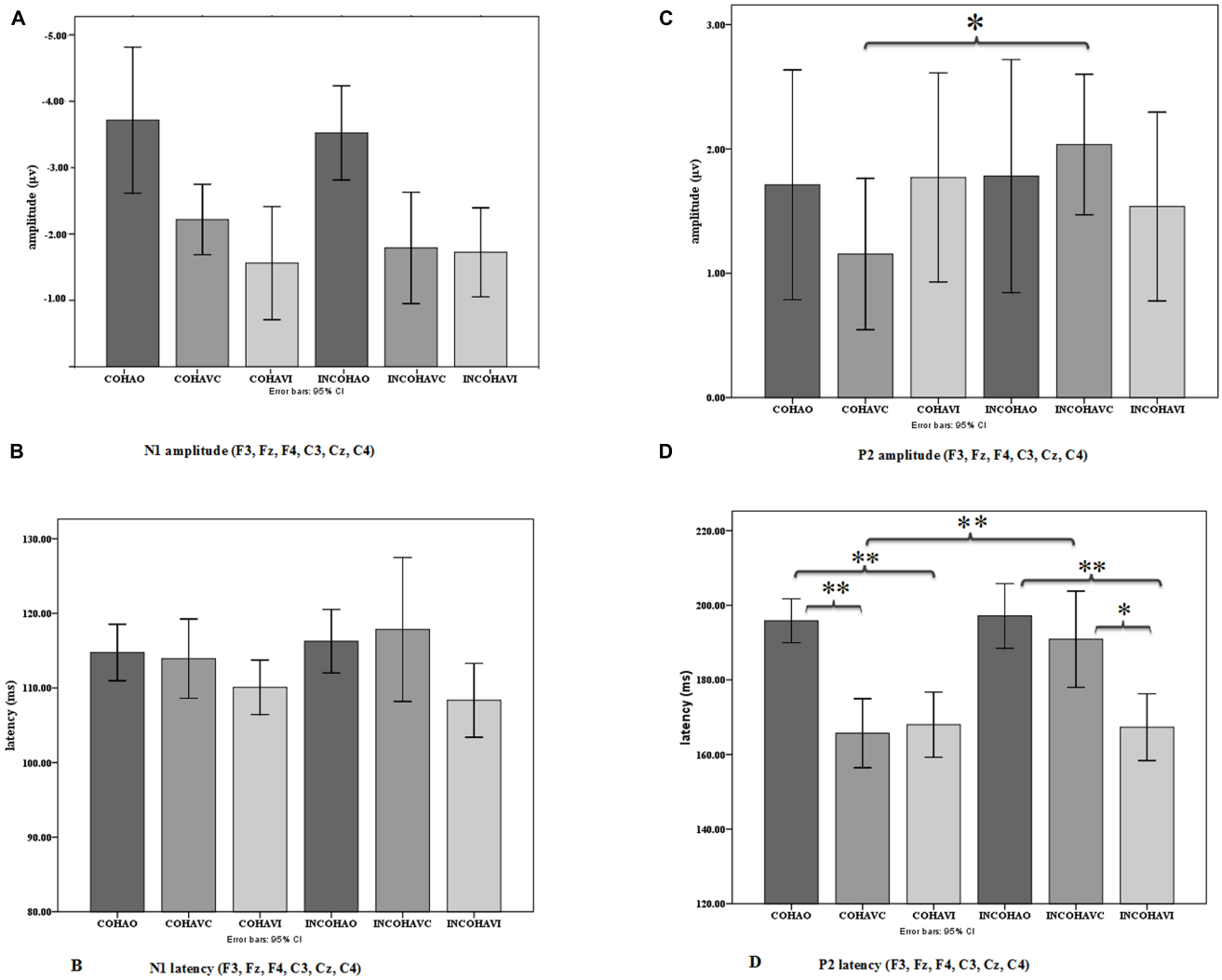


FIGURE 8 | Mean N1 (A) and P2 (C) amplitude and mean N1 (B) and P2 (D) latency averaged over the six electrodes in the three conditions (AO, AVC, and AVI) and the two contexts (coherent and incoherent). Error bars represent standard errors of the mean. $*p = 0.05$; $**p = 0.005$.

Scalp topographies and the potential role of a contamination from visual areas (see Figures 9A–D)

To assess potential contamination of the previous responses by visually driven responses from the visual cortex, we analyzed scalp topographies in the N1-P2 time periods in various conditions. Firstly we assessed whether visual areas could intervene in the visual modulation of N1 and P2 responses in the congruent and incongruent configurations, independently on context, by comparing the AO condition (**Figure 9A**) with either the AVC (**Figure 9B**) or the AVI (**Figure 9C**) condition (averaging responses over context, that is combining Coherent AVC and Incoherent AVC in **Figure 9B** and Coherent AVI and Incoherent AVI in **Figure 9C**).

In the N1 time period (100–150 ms) it appeared that the negative peak value was more prominent in central than in occipital electrodes (**Figure 9A**), but the decrease in N1 amplitude in central electrodes in both AVC and AVI conditions, associated with a negative amplitude in central electrodes in both AO-AVC and AO-AVI maps (**Figures 9B,C**) was accompanied by an even larger negative amplitude in occipital electrodes. This is due to a positive peak in AV conditions corresponding to the arrival of the visual response in this region. Therefore a possible contamination of the visual influence on N1 response due to occipital activity cannot be discarded at this stage.

In the P2 time period (175–225 ms), once again the positive peak was more prominent in central than in occipital electrodes (**Figure 9A**). The AO-AVC and AO-AVI scalp maps (**Figures 9B,C**) displayed positive values in central electrodes, corresponding to a decrease in P2 amplitude from AO to both AV conditions. Contrary to what happened for N1, the situation in occipital electrodes was here completely reversed: there were indeed negative values of AO-AVC and AO-AVI differences in the occipital region. Therefore, the possible contamination of visual effects on P2 by visual responses is much less likely than for N1.

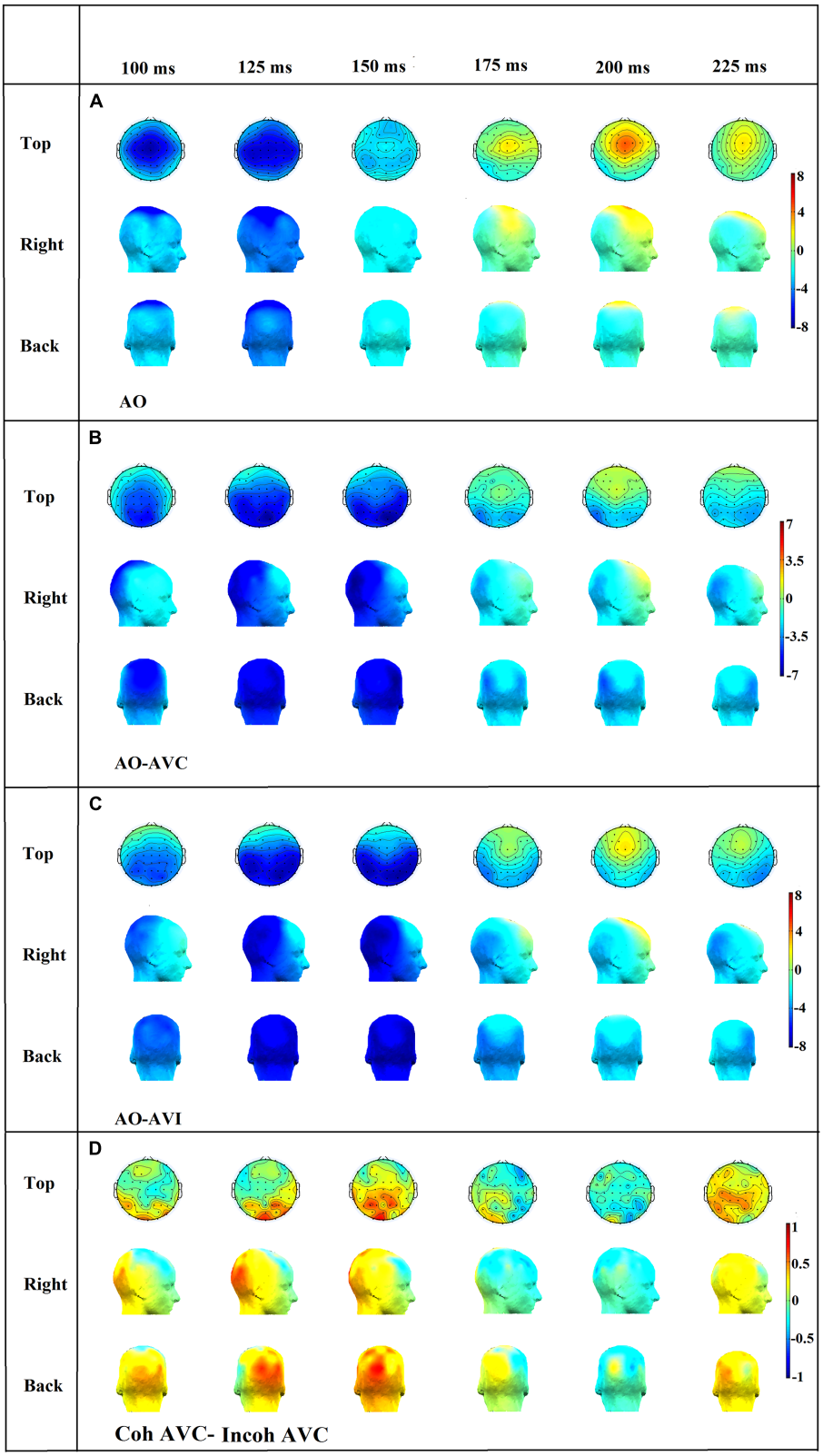


FIGURE 9 | Topographical distributions of the grand average ERPs for the AO (A), AO-AVC (B), AO-AVI (C) and Coh AVC-Incoh AVC (D) different waves in time steps of 25 ms. The range of the voltage maps varies between maps, but is always expressed in μv.

Finally, to directly assess possible contaminations on the major effect of interest that is the difference between incoherent and coherent contexts in the AVC condition, we computed scalp topographies for the difference between coherent AVC and incoherent AVI conditions (see **Figure 9D**). The differences were rather small all over these maps, and the topography differences were globally relatively noisy and make difficult any clear-cut conclusion from these topography.

Altogether, the results in the coherent context condition seem partially consistent with previous findings of EEG studies, if we assume that the Coherent context provides a condition similar to previous studies with no-context. Visual speech in the congruent AVC and incongruent AVI conditions is associated to both a significant decrease in amplitude for N1 and in latency for P2. Importantly we found a significant effect of context in the AVC condition for both amplitude and latency in P2, in line with our prediction. However, scalp topographies raise a number of questions and doubts on the possibility to unambiguously interpret these data, in the absence of a visual-only condition. We will now discuss these results in relation with both previous EEG studies on AV speech perception and with our own assumptions on AV binding.

DISCUSSION

Before discussing these results it is necessary to consider one important potential limitation of the present findings. Testing cross-modal interactions usually involves determining whether the observed response in the bimodal condition differs from the sum of those observed in the unimodal conditions (e.g., $AV \neq A + V$). In the present study, as previously noted, the visual-alone condition was not obtained because of time limitation. Although direct comparison between AV and auditory conditions performed in previous EEG studies on AV speech integration have provided fully coherent results with other studies using an additive model (see van Wassenhove et al., 2005; Pilling, 2009; Treille et al., 2014a,b), this limitation is important, and will lead to a specific component of our discussion.

COMPARISON OF THE COHERENT CONTEXT CONDITIONS WITH PREVIOUS EEG STUDIES

A preliminary objective of the study was to replicate the results of previous EEG studies on N1/P2 in coherent context (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007, 2012; Pilling, 2009; Vroomen and Stekelenburg, 2010; Baart et al., 2014; Knowland et al., 2014; Treille et al., 2014a,b). Concerning AV congruent stimuli AVC, our data are partially in line with previous studies. For the N1 component, we obtained an amplitude reduction in AVC compared to AO, as in previous studies (**Figure 8A**), though this amplitude reduction was not accompanied by a latency reduction (**Figure 8B**), contrary to previous studies. In the P2 component, the decrease in amplitude and latency (**Figures 8C,D**) from AO to AVC is also in line with previous studies (e.g., van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2009; Vroomen and Stekelenburg, 2010; Knowland et al., 2014). Concerning AV incongruent ("McGurk") stimuli AVI, there was an amplitude reduction compared to the AO condition for N1 (**Figure 8A**) and the two peaks also occurred earlier than in the AO condition, not significantly in N1 (**Figure 8B**) but significantly in P2 (**Figure 8D**). Here, the output of previous studies is more contrasted. As a matter of fact, the N1 amplitude and latency values for incongruent stimuli are not available in the van Wassenhove et al. (2005) study, whereas in the studies by Stekelenburg and Vroomen (2007) and Baart et al. (2014) there is no difference between incongruent and congruent conditions on both amplitude and latency. However, the results for P2 are not consistent with the previous studies that compared congruent and incongruent stimuli, e.g., in the study by Stekelenburg and Vroomen (2007) there is an effect of incongruent stimuli on amplitude but no effect on latency whereas in the study by van Wassenhove et al. (2005) there is no amplitude effect but a latency effect. On the contrary, the recent study by Knowland et al. (2014) is in line with the present findings in the incongruent condition for N1 and P2 amplitude, even though the stimulus for incongruency differs from the present study. Of course, some of these differences could also be due to various methodological differences in the analyses, including in the present case the specific choice to systematically keep the first peak in the P2 region in the case of double peaks responses, which occur for many subjects (see Analyses).

COMPARISON OF THE COHERENT AND INCOHERENT CONTEXT CONDITIONS IN THE PRESENT STUDY

The primary objective of the study was to test the possible role of an incoherent context supposed to lead to unbinding (as robustly displayed by behavioral data in Nahorna et al., 2012, 2013) and hence decrease the effects of the visual input on N1/P2 latency and amplitude.

We obtained no effect of context, either alone or in interaction with modality, for both N1 amplitude and latency (**Figures 8A,B**). However, we obtained a significant effect of context for P2, alone for latency, and in interaction with modality for both latency and amplitude. *Post hoc* tests showed that these effects could be due to a suppression of the decrease in amplitude and latency from AO to AVC when the context is incoherent (**Figures 8C,D**).

The fact that there is an effect of context for P2 but not for N1 is coherent with the view that these components could reflect different processing stages, AV effects on N1 possibly being not speech specific and only driven by visual anticipation independently on AV phonetic congruence, while P2 would be speech specific, content dependent and modulated by AV coherence (Stekelenburg and Vroomen, 2007; Baart et al., 2014). In summary, the visual modality would produce a decrease in N1 amplitude and possibly latency because of visual anticipation, independently on target congruence and context coherence. A congruent visual input (AVC) would lead to a decrease in P2 amplitude and latency in the coherent context because of visual predictability and AV speech-specific binding. This effect would be suppressed by incoherent context because of unbinding due to incoherence.

As for AVI stimuli, there was no context effect, both in behavioral and EEG results. Actually, it appears that there is almost no AV integration in the present study for incongruent McGurk stimuli (as shown by behavioral data), which likely explains the lack of a role of context on EEG for these stimuli. The discrepancy

in behavioral data with previous experiments by Nahorna et al. (2012, 2013) likely comes from differences in the nature and number of stimuli. The studies by Nahorna et al. (2012, 2013) involved voiced stimuli "ba," "da," and "ga" whereas in the present study the EEG requirement to avoid prevoicing, classical in the French language, forced us to select unvoiced stimuli "pa," "ta," and "ka." More importantly, the previous studies were based on a larger level of unpredictability, the subjects did not know when the targets would happen in the films, and the coherent and incoherent contexts were systematically mixed. In the present study, because of the constraints in the EEG paradigm, there were no temporal uncertainty of the time when the target occurred, and the AV material was highly restricted, with only 10 different stimuli altogether (five different targets and two different contexts). A perspective would hence be to use more variable stimuli in a further experiment.

The difference between AO and AVI conditions in P2 latency and amplitude could be related to the fact that the subjects detect an AV incongruence. Indeed, behavioral data in Nahorna et al. (2012, 2013) consistently display an increase in response times for McGurk stimuli compared with congruent stimuli, independently on context, and this was interpreted by the authors as suggesting that subjects detected the local incongruence independently on binding *per se*, while binding would modulate the final decision. In summary, AVI would produce (i) decrease in N1 amplitude and possibly latency because of visual anticipation; (ii) decrease of P2 amplitude and latency because of incongruence detection; (iii) but no integration *per se*, as displayed by behavioral data, and hence no modulation by context and binding/unbinding mechanisms.

At this stage, and keeping for a while this global interpretation compatible with the "binding" hypothesis, it is possible to come back to the two-stage AV fusion process (**Figure 1**). The present EEG data add some information about the way coherence could be computed for congruent stimuli. If indeed the P2 AV modulation in amplitude and latency is related to the binding mechanism as supposed by, e.g., Baart et al. (2014), then the evaluation of coherence, supposed by Nahorna et al. (2012, 2013) to take place in the context period before the target, should apply for both congruent and incongruent stimuli. Actually, modulation of binding by context has been shown in behavioral data on incongruent stimuli in previous studies, and in P2 data on congruent stimuli in the present study. Altogether, this suggests that the two-stage process described in **Figure 1** could operate, at least in part, prior to P2. These findings will have to be confirmed by future EEG experiments on more variable stimuli able to provide P2 modulation for both congruent and incongruent stimuli, and possibly in other kinds of attentional processes.

POSSIBLE CONTAMINATION BY VISUAL AREAS AND SUGGESTIONS FOR FUTURE STUDIES

A crucial limitation of the present work is the lack of a visual-only condition. We consider that this was a necessary evil in such a preliminary study, since it was the only way to be able to assess both congruent and incongruent targets in coherent vs. incoherent contexts. But this might have resulted in possible contamination effects from visual regions that we will discuss now.

Firstly, contamination could be due to visual context. This is, however, rather unlikely considering that the different contexts finish 1000 ms before the target. We systematically compared results obtained with two baseline conditions, one far from the end of the context (–100 to 0 ms) and the other one closer (–500 to –400 ms). It appeared that this baseline change did not change the current results in any crucial way, either in whole graphs or statistical analysis, which suggests that the fluctuations in ERP responses before the apparition of the auditory stimulus at 0 ms do not intervene much in the further analysis of AV interactions on N1 and P2.

It is more likely that contamination effects could be due to visual responses to the visual component of the target. This appears particularly likely in the N1 time period, where scalp maps in the AO-AVC and AO-AVI conditions (**Figure 9**) display larger negative values in occipital areas than in central electrodes. Therefore, it cannot be ruled out that (some unknown) part of the visual modulation of the auditory response could be due to propagation of visual responses from the occipital region.

In the P2 time period this is much less likely, considering that the pattern of responses is now completely inverse between central and occipital electrodes, with a decrease of P2 amplitude from AO to AVC or AVI in the first ones, and an increase in the second ones. However, the pattern of scalp difference between coherent and incoherent AVC conditions is complex and fuzzy, and the amplitude differences between conditions are small. Therefore, we cannot discard the possibility that the modulation of P2 response in the incoherent compared with coherent context is due to propagation of the visual activity – though we must remind that in these two conditions, the visual response actually corresponds to exactly the same visual input, which makes the "visual propagation" hypothesis more unlikely.

Altogether our interpretation of the observed results is that (1) the pattern of EEG responses we obtained in the N1-P2 time periods is compatible with classical visual effects on the auditory response in this pattern of time, and with a possible modulation of these effects by AV context, in line with our assumptions on AV binding; (2) however, the lack of a visual-only condition impedes to firmly discard other interpretations considering contamination from visual regions due to responses to the visual component of the stimulus; and (3) this suggests that more experiments using the same kind of paradigm with AV context, incorporating visual-only conditions to enable better control of the visual effects are needed to assess the possibility to exhibit electrophysiological correlates of the binding/unbinding mechanism in the human brain.

CONCLUSION

We displayed a new paradigm for ERP AV studies based on the role of context. We presented data about modulation of the auditory response in the N1-P2 time periods due to the visual input, both in the target and context portions of the stimulus. We proposed a possible interpretation of the modulations of the N1 and P2 components, associated to (1) a classical visual modulation generally associated with predictive mechanisms (see e.g., van Wassenhove et al., 2005) and (2) possible modifications of this effect due to incoherent context, in the framework of the two-stage "binding and fusion" model proposed by Nahorna et al. (2012). However, we also discussed in detail a concurrent interpretation only based on

the contamination by visual responses in the visual regions, due to the impossibility in the present study to incorporate a visual-only condition.

The search for electrophysiological correlates of attentional processes possibly modifying AV interactions is an important challenge for research on AV speech perception (see e.g., the recent study by Alsius et al. (2014) measuring the effect of attentional load on AV speech perception using N1 and P2 responses as cues just as in the present study). We suggest that binding associated with context should be integrated in general descriptions of AV modulations of the N1 and P2 components of auditory ERP responses to speech stimuli, in relation with general and speech specific effects and the role of attention.

REFERENCES

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S. S., and Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Front. Psychol.* 5:727. doi: 10.3389/fpsyg.2014.00727

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., and Soto-Faraco, S. S. (2007). Attention to touch weakens audiovisual speech integration. *Exp. Brain Res.* 183, 399–404. doi: 10.1007/s00221-007-1110-1

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Baart, M., Stekelenburg, J. J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 65, 115–211. doi: 10.1016/j.neuropsychologia.2013.11.011

Benoît, C., Mohamadi T., and Kandel S. (1994). Effects of phonetic context on audio-visual intelligibility of French speech in noise. *J. Speech Hear. Res.* 37, 1195–1203. doi: 10.1044/jshr.3705.1195

Bernstein, L. E., Auer, E. T., and Moore, J. K. (2004a). "Audiovisual speech binding: convergence or association?," in *The Handbook of Multisensory Processes*, eds G. A. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: The MIT Press), 203–224.

Bernstein, L. E., Auer, E. T., and Takayanagi, S. (2004b). Auditory speech detection in noise enhanced by lipreading. *Speech Commun.* 44, 5–18. doi: 10.1016/j.specom.2004.10.011

Berthommier, F. (2004). A phonetically neutral model of the low-level audiovisual interaction. *Speech Commun.* 44, 31–41. doi: 10.1016/j.specom.2004.10.003

Bertoli, S., Probst, R., and Bodmer, D. (2011). Late auditory evoked potentials in elderly long-term hearing-aid users with unilateral or bilateral fittings. *Hear. Res.* 280, 58–69. doi: 10.1016/j.heares.2011.04.013

Besle, J., Fort, A., Delpuech, C., and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Buchan, J. N., and Munhall, K. G. (2012). The effect of a concurrent cognitive load task and temporal offsets on the integration of auditory and visual speech information. *Seeing Perceiving* 25, 87–106. doi: 10.1163/187847611X620937

Ceponiene, R., Westerfield, M., Torki, M., and Townsend, J. (2008). Modality-specificity of sensory aging in vision and audition: evidence from event-related potentials. *Brain Res.* 1215, 53–68. doi: 10.1016/j.brainres.2008.02.010

Colin, C., Radeau-Loicq, M., and Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *Eur. J. Cogn. Psychol.* 17, 541–560. doi: 10.1080/09541440440000168

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *J. Speech Hear. Res.* 14, 496–512. doi: 10.1044/jshr.1403.496

Grant, K. W., and Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208. doi: 10.1121/1.1288668

Hyde, M. (1997). The N1 response and its applications. *Audiol. Neurootol.* 26, 281–307. doi: 10.1159/000259253

Kim, J., and Davis, C. (2004). Investigating the audio-visual detection advantage. *Speech Commun.* 44, 19–30. doi: 10.1016/j.specom.2004.09.008

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Knowland, V. C. P., Mercure, E., Karmiloff-Smith, A., Dick, F., and Thomas, M. S. C. (2014). Audio-visual speech perception: a developmental ERP investigation. *Dev. Sci.* 17, 110–124. doi: 10.1111/desc.12098

Lallouache, M. T. (1990). "Un poste 'visage-parole.' Acquisition et traitement de contours labiaux (A 'face-speech' workstation. Acquisition and processing of labial contours)," in *Proceedings of the eighteenth Journées d'Etudes sur la Parole*, Montréal, QC.

Massaro, D. W. (1987). *Speech Perception by Ear and Eye.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 265, 746–748. doi: 10.1038/264746a0

Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., and Laurienti, P. J. (2008). Modality-specific selective attention attenuates multisensory integration. *Exp. Brain Res.* 184, 39–52. doi: 10.1007/s00221-007-1080-3

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Näätänen, R., and Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychol. Bull.* 6, 826–859. doi: 10.1037//0033-2909.125.6.826

Nahorna, O., Berthommier, F., and Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* 132, 1061–1077. doi: 10.1121/1.4728187

Nahorna, O., Ganesh, A. C., Berthommier, F., and Schwartz, J. L. (2013). "Modulating fusion in the McGurk effect by binding processes and contextual noise," in *Proceedings of the 12th international conference on auditory-visual speech processing*, Annecy.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., and Spence, C. (2005). Exposure to asynchronous audiovisual speech increases the temporal window for audiovisual integration of non-speech stimuli. *Cogn. Brain Res.* 25, 499–507. doi: 10.1016/j.cogbrainres.2005.07.009

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J. Speech Lang. Hear. Res.* 52, 1073–1081. doi: 10.1044/1092-4388(2009/07-0276)

Ponton, C. W., Don, M., Eggermont, J. J., Waring, M. D., and Masuda, A. (1996). Maturation of human cortical auditory function: differences between normal-hearing children and children with cochlear implants. *Ear. Hear.* 17, 430–437. doi: 10.1097/00003446-199610000-00009

Scherg, M., and Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroenceph. Clin. Neurophysiol.* 65, 344–360. doi: 10.1016/0168-5597(86)90014-6

Schwartz, J. L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition* 93, B69–B78. doi: 10.1016/j.cognition.2004.01.006

Schwartz, J. L., Robert-Ribes, J., and Escudier, P. (1998). "Ten years after Summerfield. A taxonomy of models for audiovisual fusion in speech perception," in *Hearing by Eye II. Perspectives and Directions in Research on Audiovisual Aspects of Language Processing*, eds R. Campbell, B. Dodd, and D. Burnham (Hove: Psychology Press), 85–108.

Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* 92, B13–B23. doi: 10.1016/j.cognition.2003.10.005

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.* 19, 1964–1973. doi: 10.1162/jocn.2007.19.12.1964

Stekelenburg, J. J., and Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Front. Integr. Neurosci.* 6:26. doi: 10.3389/fnint.2012.00026

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading*, eds B. Dodd and R. Campbell (New York: Lawrence Erlbaum), 3–51.

Tiippana, K., Andersen, T. S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

Treille, A., Cordeboeuf, C., Vilain, C., and Sato, M. (2014a). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. *Neuropsychologia* 57, 71–77. doi: 10.1016/j.neuropsychologia.2014.02.004

Treille, A., Vilain, C., and Sato, M. (2014b). The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. *Front. Psychol.* 5:420. doi: 10.3389/fpsyg.2014.00420

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596. doi: 10.1162/jocn.2009.21308

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 March 2014; accepted: 03 November 2014; published online: 26 November 2014.

Citation: Ganesh AC, Berthommier F, Vilain C, Sato M and Schwartz J-L (2014) A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. Front. Psychol. ***5****:1340. doi: 10.3389/fpsyg.2014.01340*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 16 July 2014
doi: 10.3389/fpsyg.2014.00698

Talker variability in audio-visual speech perception

Shannon L. M. Heald* and Howard C. Nusbaum

Department of Psychology, The University of Chicago, Chicago, IL, USA

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Yang Zhang, University of Minnesota, USA
Pia Knoeferle, Bielefeld University, Germany

****Correspondence:***
Shannon L. M. Heald, Department of Psychology, The University of Chicago, 5848 South University Avenue – B406, Chicago, IL 60637, USA
e-mail: smbowdre@uchicago.edu

A change in talker is a change in the context for the phonetic interpretation of acoustic patterns of speech. Different talkers have different mappings between acoustic patterns and phonetic categories and listeners need to adapt to these differences. Despite this complexity, listeners are adept at comprehending speech in multiple-talker contexts, albeit at a slight but measurable performance cost (e.g., slower recognition). So far, this talker variability cost has been demonstrated only in audio-only speech. Other research in single-talker contexts have shown, however, that when listeners are able to see a talker's face, speech recognition is improved under adverse listening (e.g., noise or distortion) conditions that can increase uncertainty in the mapping between acoustic patterns and phonetic categories. Does seeing a talker's face reduce the cost of word recognition in multiple-talker contexts? We used a speeded word-monitoring task in which listeners make quick judgments about target word recognition in single- and multiple-talker contexts. Results show faster recognition performance in single-talker conditions compared to multiple-talker conditions for both audio-only and audio-visual speech. However, recognition time in a multiple-talker context was slower in the audio-visual condition compared to audio-only condition. These results suggest that seeing a talker's face during speech perception may slow recognition by increasing the importance of talker identification, signaling to the listener a change in talker has occurred.

Keywords: talker normalization, talker variability, audio-visual speech perception, multisensory integration, speech perception

INTRODUCTION

In perceiving speech, we listen in order to understand what someone is saying as well as to understand who is saying it. Although the message changes more often in a conversation, there can also be changes between speakers that are important for the listener to recognize. A change in talker can pose a perceptual challenge to a listener due to an increase in the variability of the way acoustic patterns map on to phonetic categories – a problem of talker variability. For different talkers, a given acoustic pattern may correspond to different phonemes, while conversely, a given phoneme may be represented by different acoustic patterns across different talkers (Peterson and Barney, 1952; Liberman et al., 1967; Dorman et al., 1977). For this reason, the speaker provides an important context to determine how acoustic patterns map on to phonetic categories (cf. Nusbaum and Magnuson, 1997). Additionally, a change in talker may be important to recognize given that a listener's interpretation of a message may depend not just on the speech style of a speaker, but on the attributions about who the speaker is as well (Thakerar and Giles, 1981). For example, indirect requests are understood in the context of a speaker's status (Holtgraves, 1994). More directly relevant to speech perception however, a listener's belief about the social group to which a speaker belongs can significantly alter the perceived intelligibility of a speaker's speech (Rubin, 1992). Additionally, dialect (Niedzielski, 1999) and gender (Johnson et al., 1999) expectations can meaningfully alter vowel perception, highlighting that social knowledge about a speaker can affect the relatively low-level perceptual processing of a speaker's message, much in the same way that knowledge of vocal tract information can (Ladefoged and Broadbent, 1957; although see Huang and Holt, 2012 for an auditory explanation of the mechanism that could underlie this).

In general there have been two broad views regarding how talker information is recognized. One account, called "talker normalization" (Nearey, 1989; Nusbaum and Magnuson, 1997), suggests that listeners use talker information to calibrate or frame the interpretation of a given message in order to overcome the considerable amount of uncertainty (e.g., acoustic variability, reference resolution, etc.) that arises from talker differences. This view has emerged from an attempt to address the lack of invariance problem through the use of talker-specific information either derived from the context of prior speech (Joos, 1948; Ladefoged and Broadbent, 1957; Gerstman, 1968) or cues within the utterance (e.g., Syrdal and Gopal, 1986). The sufficiency of such models has been demonstrated for vowel perception (e.g., Gerstman, 1968; Syrdal and Gopal, 1986) for both types of approaches. Further, perceptual evidence has come from demonstrations of better recognition for speech from a single-talker compared to speech from different talkers (e.g., Creelman, 1957; Nearey, 1989) and that specific acoustic information can aid in normalizing talker differences (e.g., Nusbaum and Morin, 1992; Barreda and Nearey, 2012).

An alternative view regarding how talker information is recognized suggests that talker information is not used in direct service of message understanding but for source understanding. This view treats the identification of the talker as separate from

the process of message comprehension (Pisoni, 1997; Goldinger, 1998). Traditionally, speech perception has been described as a process whereby linguistic units (e.g., phonemes, words) are abstracted away from the detailed acoustic information that is putatively not phonetically relevant. The idea that acoustic information about a talker might be viewed as noise in relation to the canonical linguistic units upon which speech perception relies, has led to the assumption that talker information is lost during this process (e.g., Joos, 1948; Summerfield and Haggard, 1973; Halle, 1985; McLennan and Luce, 2005)[1]. However, the need for preserving talker-specific information for other perceptual goals (Thakerar and Giles, 1981; Holtgraves, 1994), along with evidence suggesting that the perceptual learning of speech is talker-specific (Goldinger et al., 1991; Schacter, 1992; Pisoni, 1993; Nygaard et al., 1994) prompted researchers to adopt a talker-specific view of speech perception.

In the talker-specific view, auditory representations of utterances are putatively represented in a more veridical fashion. As such, both the indexical source auditory information is maintained along with any phonetically relevant auditory information (e.g., Goldinger, 1998). While this view does separately preserve talker-specific auditory information such as fundamental frequency within the auditory-trace, the model has no implications for the representation or processing of other aspects of talker information such as knowledge about the social group of the talker, the dialect of the talker, or the gender of the talker. Further, the echoic encoding account does not explain how talker-specific information that is not in the acoustic channel affects speech processing, as it focuses on the memory representation of auditory patterns.

A number of studies have demonstrated that in a variety of learning situations, variability is important in developing robust perceptual categories that can benefit recognition in diverse listening conditions. In particular, variability in talker has been shown to benefit the long-term memory representations of speech that can facilitate recognition when there is noise or degraded signal or in learning a foreign contrast (Logan et al., 1991; Nygaard et al., 1994; Zhang et al., 2009). However, these studies tend to focus on the benefits of variability in the learning process during which phonetic representations or lexical representations are formed for use in recognition. But beyond this variability in the process of learning speech representations, there is also variability in the moment when one talker stops speaking and another starts. This kind of variability has a short-term effect of slowing recognition, shifting attention to different acoustic properties and increasing activity consistent with an attentionally demanding process (Mullennix and Pisoni, 1990; Nusbaum and Morin, 1992; Wong et al., 2004; Magnuson and Nusbaum, 2007). The difference in these two kinds of situations is not simply that the goal of one set of studies is learning (learning a talker or phonological or lexical forms) vs. speeded recognition, but also that the studies of learning are not designed to evaluate the nature of processing that occurs in the first 10 ms of encountering a new talker but instead focus on the nature of the representations ultimately developed. However, as has been discussed for many decades from Ladefoged and Broadbent (1957) to Barreda and Nearey (2012), variability in the mapping between acoustic patterns and linguistic categories differs across talkers and this variability has been shown to elicit worse performance across a number of measures [slower response times (RTs), lower hit rate, or higher false alarm rate; Wong et al., 2004; Magnuson and Nusbaum, 2007]. Further, the evidence that these performance costs are not mitigated by familiarizing listeners with the talkers (Magnuson et al., 1994) suggests that there is a clear separation between talker variability effects on the short-term accommodation to speech and learning effects in a multi-talker context.

While familiarity with a talker does not appear to influence the talker variability effect found in the short-term accommodation to speech, it remains unclear whether non-acoustic information about a talker can moderate the effect of talker variability. Much of the research regarding talker variability effects has examined the notable acoustic variability found in a multiple-talker context. However, a multiple-talker context can produce variability in other sensory channels (beyond the acoustic), which could impair talker identification and message comprehension. Given that conversations can take place among several interlocutors in a face-to-face context, it is reasonable to ask how the presence of face information affects speech perception when the talker changes. If watching a talking face provides cues for both talker identification and message comprehension there are two potential effects. One possibility is that seeing a new talker will slow recognition, as it will prompt the listener to enter into an attention-demanding (Nusbaum and Morin, 1992; Wong et al., 2004) process by which the speech of the new talker is perceptually normalized (Nearey, 1989; Nusbaum and Magnuson, 1997). Conversely, the presence of face information may speed up recognition by providing a converging source of phonetic information through visemes that allows the listener to achieve faster and/or more accurate word recognition (Sumby and Pollack, 1954; Summerfield, 1987; Massaro and Cohen, 1995; Rosenblum et al., 1996; Lachs et al., 2001).

Previous research has demonstrated that a person's face is an important source of information about social category membership, which can also influence speech perception. As noted already, the subjectively rated intelligibility of the same speech signal is different depending on whether the speech is accompanied by pictures of putative speakers from different racial groups (Rubin, 1992). Similarly, the classification of vowels can be changed by seeing a different gendered face presented falsely as the speaker (Johnson et al., 1999). In both cases, participants simply viewed static photographs that identified the speaker. Given human face expertise (e.g., Diamond and Carey, 1986; Gauthier and Nelson, 2001), observers are very accurate in recognizing faces (Bahrick et al., 1975), even more so than in recognizing voices (Read and Craik, 1995; Olsson et al., 1998; Wilding and Cook, 2000). Thus, the presence of visual face information provides an ecologically reliable cue about speaker identity. Work by Magnuson and Nusbaum (2007) has demonstrated that the effect of talker

[1]Although, it is possible that talker information, even under a talker normalization rubric, is preserved in parallel representational structures for other listening goals (e.g., Hasson et al., 2007).

variability can be mediated entirely by expectations the listener holds regarding talker differences. This study showed that when an acoustic difference (a small F0 difference) was attributed to normal production variability of a single-talker, variation in F0 did not slow recognition down any more than a constant F0. However, when the identical acoustic difference was interpreted (based on prior expectation) as a talker difference, the same F0 variability led to slower recognition compared to a condition with a constant F0. This demonstrates that it is not the acoustic variability that slows recognition but the knowledge of what that variability means. Seeing a face change provides similar knowledge to listeners, as it signals to the listeners that a change in talker has indeed occurred. Therefore, it is reasonable that visual face information may act to signal a change in talker and therefore the need to calibrate perception through normalization.

While there is evidence that a still photograph can give clear information about the identity of a speaker, a video of the speaker's face provides additional information, as a talking face can additionally show visible articulatory gestures. For example, the intelligibility of speech in noise (Sumby and Pollack, 1954) as well as speech heard through cochlear implants (Goh et al., 2001; Lachs et al., 2001) is significantly improved by additionally seeing a speaker talk. However, there is clear evidence that the visual information of mouth movements is not simply redundant with the speech signal. The McGurk and MacDonald (1976) effect clearly demonstrates that independent articulatory information can be visually gleaned and integrated with speech signals during perception. To engender the McGurk and MacDonald (1976) effect, a participant is shown a video of a mouth producing one place of articulation (e.g., /ka/) while hearing acoustic information corresponding to a different place of articulation at the same time (e.g., /pa/). This presentation combination results in the perception of a third illusory place of articulation (e.g., /ta/). Indeed, using neuroimaging during the presentation of McGurk stimuli, Skipper et al. (2007) demonstrated that the pattern of brain activity in the supramarginal gyrus starts out consistent with the acoustic information (e.g., /pa/) but changes over time to be consistent with the final percept (i.e., /ta/), whereas brain activity in the middle occipital gyrus starts out consistent with the visual mouth movements (e.g., /ka/) but ends up responding with a pattern consistent with the final percept. However, the ventral premotor region starts out coding the perceptual category and maintains that activity pattern. The illusion along with the neuroimaging data suggests that different sensory systems initially code different sources of perceptual information about speech in interaction with divergent information represented in the motor system. If seeing mouth movements improves recognition performance as shown behaviorally by recruiting premotor cortex and increasing superior temporal activity (Skipper et al., 2005, 2007), it is possible that slower recognition and/or worse accuracy associated with a change in talker might be ameliorated if not eliminated, given that seeing mouth movements may provide additional information such as visemes that could be used to limit or constrain phonetic interpretation from the acoustic channel.

Thus seeing a talker can visually provide both message-relevant and source-relevant information, just as the acoustic pattern of an utterance does. On the one hand, a face can convey clear talker identity information to an observer, which can be important when listening to speech because it may signal a change in talker and the need to calibrate perception through normalization. On the other hand, mouth movements can additionally convey articulatory information that may help constrain acoustic variability. Although Olsson et al. (1998) have shown that speech is a much more effective cue to message content than mouth movements, Rosenblum et al. (1996) have demonstrated that even with the low accuracy of lip reading, this information significantly boosts the recognition of spoken words in noise. Given these two different possibilities for the way that visual information is used by listeners, it is unclear how seeing talkers would affect speech recognition when there is talker variability. Visual talker information could act as a strong signal of talker change (thereby requiring more perceptual analysis of the face and speech) ultimately slowing speech recognition. Conversely, the presence of a face could speed up recognition through the provision of concurrent viseme information.

The present study was carried out to address how seeing a talker would influence speech recognition in a multiple-talker context. Listeners performed a speeded word recognition task, listening for spoken words that were designated as a target. Targets differed in several phonemes from other targets and distracters to ensure that recognition did not depend on a single phonetic contrast. Listeners were required to respond every time they recognized a target. On each trial, four occurrences of a target word were presented randomly in a sequence along with 12 randomly selected distracters. On single-talker trials, one talker produced all the target and distracter speech, while in multiple-talker trials, multiple-talkers produced both targets and distracters. In the present study, one group (half of the participants) was presented with only the acoustic speech signal. This portion of the study replicates the design of previous, audio-only talker variability studies using speeded target detection (e.g., Nusbaum and Morin, 1992; Wong et al., 2004; Magnuson and Nusbaum, 2007). A second group (half of the participants) was presented with audio-visual speech in which the listener could see and hear the talker producing the utterance. Previous, audio-only, talker variability studies have demonstrated better performance (fast reaction times, higher hit rate, or lower false alarm rate) for single-talker trials compared to multiple-talker trials (Wong et al., 2004; Magnuson and Nusbaum, 2007).

There are two possible predictions regarding the way that seeing a talker will influence speech recognition speed in the present study. If seeing a talker's mouth movements provides viseme information to reduce acoustic-phonetic uncertainty, then audio-visual speech will have better performance than audio-only speech, independent of how much talker variability is present. Further, viseme information present when seeing a talker could also reduce, if not eliminate the poorer recognition performance associated with talker variability. Performance in the multiple-talker condition could be improved if viseme information constrains the one-to-many mapping of acoustic segments onto phonetic categories. If

this is the case then recognition performance for single-talker trials should not significantly differ from recognition performance for multiple-talker trials in the audio-visual condition. Indeed, the poorer performance found in multiple-talker trials in audio-only studies may be an artifact of the "unnatural" (in the context of evolution) situation of hearing speech without seeing the talkers.

Another possible prediction however, is that seeing talkers may be a much more powerful signal of talker identity than simply hearing speech. If so, then seeing talkers might result in even poorer performance than has been found in multiple-talker trials compared to single-talker trials, if the face acts as a cue for listeners to enter into a talker normalization process. If this is the case then both audio-only and audio-visual speech should both show poorer performance in the multiple-talker condition when compared to single-talker condition. Further, if the presence of the face does act as a more effective cue to talker change, then the multiple-talker condition might show even poorer performance in audio-visual condition compared to audio-only condition. This would be the case if audio-only speech is a less effective cue to talker change than audio-visual speech and as such, results in producing more occurrences of talker normalization in the audio-visual condition. As poorer performance could manifest as an increase in reaction time, a decrease in hit rate, an increase in false alarm rate or a drop in d-prime, every participant's average RT, hit rate, false alarm rate, and d-prime were measured for each condition.

MATERIALS AND METHODS

PARTICIPANTS

Forty-six participants (31 female) were recruited from the University of Chicago undergraduate community and were between 18 and 26 years of age. One participant was dropped from analysis due to a technical problem in collecting data, and a further participant was excluded from analysis due to reported excessive fatigue (her overall accuracy was 79%). Both of the excluded participants were female. All of the participants were native speakers of American English, with no history of hearing, speech, or vision disorders reported. Participants were compensated with course credit and were debriefed upon the conclusion of the experimental session. Additionally, informed consent, using a form approved by the University of Chicago Institutional Review Board, was obtained from all subjects.

STIMULI

The stimuli consisted of audio-visual and audio-only versions of the same recordings of words, produced by three talkers, as different groups of listeners performed speeded word recognition for different pairs of speakers. Specifically, half of the participants performed the speeded word recognition with speech from two male talkers (Talker CL and Talker SH), while the other half of participants performed the speeded word recognition with speech from a male and a female talker (the same stimuli by Talker SH were used again, and Talker CL was replaced by Talker SK, a female talker). This was done so as to ensure that any differences we found were not due to a particular pair of speakers. The words used as stimuli were selected from the Harvard phonetic-balanced word list (IEEE Subcommittee on Subjective Measurements, 1969). We selected the words used by Magnuson and Nusbaum (2007), namely: "ball," "bluff," "cad," "cave," "cling," "depth," "dime," "done," "gnash," "greet," "jaw," "jolt," "lash," "knife," "park," "priest," "reek," "romp," and "tile." Of these 19 words, "ball," "cave," "done," and "tile" were used as target words. The stimuli were produced by all three speakers in front of a neutral green screen. The video recording was made with a Canon GL-1 digital camcorder. The visual portion of the stimuli consisted of the speaker's face directly facing the camera. The size of each talker's face was equalized across all of that talker's stimuli. Additionally, the relative differences in face size were maintained between the two speakers.

High-quality sound recordings (32 kHz, 16 bit) were simultaneously recorded along with the video using an Alesis ML-9600 sound recorder. The high-quality sound recordings were then used to replace the original soundtrack from the audio-visual recording using Finalcut Pro. The audio component of all the stimuli were RMS normalized to an average of 57.2 dB SPL. The duration of each word (from sound onset to sound offset) was measured, and the durations of words (both in terms of video and sound) produced by Talker CL and Talker SK were shortened to match the duration of each corresponding word produced by Talker SH as Talker SH had the shortest durations. Duration changes for the sound portion were accomplished by applying the PSOLA algorithm in Praat (Boersma, 2001). PSOLA was also applied to the stimuli produced by Talker SH with the speed factor of 1, as a control. Duration changes for the video portion were accomplished by altering the speed of the video in Finalcut Pro. Given that duration changes were identical for both audio and visual aspects of the recording, the final audio-visual presentation sounded natural and was free from any asynchrony. In order for the stimuli to be short enough for use in a speeded target-monitoring task, the stimuli were edited down to a length of 666 ms. In order to keep the audio portion of the audio-visual and audio-only stimuli comparable and to match stimulus durations (AV and A) across conditions, all the stimuli were edited to begin at the start of sound onset. While previous research on the time course of audio-visual speech perception has indicated that some visual cues can precede the acoustic onset by 80–100 ms (Smeele, 1994, Unpublished Doctoral dissertation; Munhall and Vatikiotis-Bateson, 1998), a gating study by Munhall and Tohkura (1998) suggests that the visual information that precedes the acoustic onset is not necessary to see a significant contributions of visual information in speech perception. Further, pretesting indicated that the stimuli were perceived as natural productions with no unnatural changes, asynchronies, or jump-cuts perceived. As such, the audio-only stimuli were equivalent to the audio-visual stimuli, except that the video channel was stripped from the audio-visual stimuli.

PROCEDURE

The experiments consisted of a speeded target-monitoring task. Before beginning the monitoring task, participants were informed that an orthographic form of a target word would be presented before every trial and that, depending on the modality condition, a sequence of audio, or audio–video recordings of spoken words would follow. Participants were instructed to press the space bar

as quickly and as accurately as possible whenever they recognized the target word. At the beginning of each trial, a fixation cross was presented at the center of a black screen for 1 s. A blank black screen was then presented for 250 ms before the printed target word (for 1 s). Another 250 ms pause preceded the presentation of the spoken stimuli. A stream of 16 spoken words was presented for each trial; each stimulus was 666 ms, followed by a silent blank screen for 84 ms before the next stimulus was presented (total SOA 750 ms). Four word targets were pseudo-randomly placed at ordinal positions between the 1st and 16th stimuli (i.e., positions 2 to 15) such that the targets were separated by at least one distractor. On each trial, one target was chosen from the set "ball," "done," "cave," and "tile." Twelve distracter words were randomly selected from the full set of stimuli, excluding the designated target (see **Figure 1**). After one practice trial, a block of 12 test trials followed, all with either stimuli from only one speaker (the single-talker condition) or from two speakers (the multiple-talker condition). In the latter condition, the talker for each of the 16 words in a trial was randomly determined. Each possible target word appeared as the target for three trials within each of four different conditions, and the order of which target was selected for a particular trial was randomized. Each participant received all four of the talker conditions (single-talker 1 condition, single-talker 2 condition, and multiple-talkers conditions combining the two talkers). Participants received either audio-visual or audio-only stimuli depending on what modality condition to which they were assigned. Every participant's RT, hit rate, false alarm rate, and d-prime were measured. Participants were always explicitly informed (both verbally and by printed instructions) of the identity of each talker condition before they began trials in that condition.

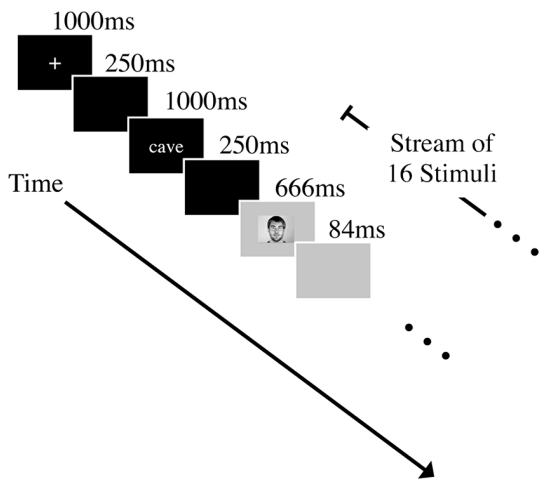


FIGURE 1 | Experimental format of an audio-visual trial. Each trial started with a fixation cross that was presented at the center of a black screen for 1000 ms. This was followed by a blank, black screen for 250 ms. Participants were then shown a printed target word (ball, done, cave, or tile) for 1000 ms. Another 250 ms pause preceded the presentation of the spoken stimuli. A stream of 16 spoken words was shown on each trial. Each stimulus was 666 ms, followed by a silent blank screen for 84 ms before the next stimulus was presented. Four word targets were pseudo-randomly placed at ordinal positions between the 1st and 16th stimuli (i.e., positions 2 to 15) such that the targets were separated by at least one distracter. Participants were instructed to press the space bar as quickly and as accurately as possible whenever they recognized the target word. Stimuli either came from only one speaker (the single-talker condition) or from two speakers (the multiple-talker condition) depending on the condition.

RESULTS

In order to examine the effect of audio-visual information on the talker variability cost, a split plot analysis of variance (ANOVA) was carried out [Talker Variability (Single-Talker vs. Multiple-Talker) × Modality of Presentation (Audio-only vs. Audio-visual), with Talker Variability as the within-subject factor and Modality of Presentation as a between-subject factor], for the dependent measures of RT, hit rate, false alarm rate, and d-prime. For the dependent measure of RT, a significant main effect of Talker Variability was found, indicating that listeners are faster to recognize speech from a single-talker (484 ms ± SEM) than from multiple-talkers [502 ms; $F(1,42) = 27.75$, $p < 0.001$]. A planned comparison indicates that the recognition time is significantly slower in the multiple-talkers trials compared to the single-talker trials in the audio-only condition [$t(21) = 1.637$, $p = 0.05$]. This replicates other audio-only talker variability work that has used this task previously (Wong et al., 2004; Magnuson and Nusbaum, 2007). There was no main effect of Modality of Presentation [$F(1,42) = 0.494$, $p = 0.48$]. A significant interaction effect of Modality × Talker Variability however, reveals that the performance cost between multiple-talker trials and single-talker trials was increased by 15 ms in the audio-visual condition (26 ms) compared to the audio-only condition [11 ms; $F(1,42) = 5.13$, $p = 0.03$]. This interaction effect, as seen in **Figure 2** is clearly driven by RT differences across modalities in the multiple-talker trials (i.e., between the audio-only multiple-talker trials and audio-visual multiple-talker trials), as there is little reaction time difference between the audio-only and audio-visual single-talker trials (mean RT in audio-only for single-talker trials was 482 ms. and mean RT in audio-visual for single-talker trials was 485 ms). Thus, it is unlikely that the interaction effect is due solely to the presence of visual information in the task, as we would have seen a similar delay in the single-talker audio-visual trials, but we did not. For this reason, the increase in RT in the audio-visual trials is likely due to extra talker information in the visual display. The same analyses were carried out using hit rate, false alarm rate, and d-prime[2] but none of these analyses yielded any significant effects or interactions (see **Table 1** for a summary of results for the DV of false alarm rate, **Table 2** for a summary of results for the DV of hit rate, and **Table 3** for a summary of results for the DV of d-prime.).

DISCUSSION

Visual information showing a speaker's mouth movements together with speech production has been shown to improve intelligibility of speech under adverse listening conditions (Sumby and Pollack, 1954; Summerfield, 1987; Massaro and Cohen,

[2]To calculate d-prime, a hit rate or false alarm rate of 1 or 0 could not be used to obtain actual z-scores (as probabilities of 1 and 0 would correspond to z-scores of ∞ and $-\infty$, respectively). For this reason, the formula $[(n * 2) \pm 1]/(t * 2)$, where n equals the total number of hits or false alarms, and t equals the total number of trials, was used as an approximation.

Table 1 | Summary of results from the split plot ANOVA [Talker Variability (Single-Talker vs. Multiple-Talkers) × Modality of Presentation (Audio-only vs. Audio-visual), with Talker Variability as a within-subject factor and Modality of Presentation as a between-subject factor] for the dependent measure of false alarm rates.

Source	*F* statistic	*p*	Estimated means (standard error)
Talker variability	0.409	0.526	0.010 (0.001) single-talker
			0.009 (0.001) multiple-talkers
Talker Variability × Modality of Presentation	2.670	0.110	0.009 (0.002) audio only single-talker
			0.010 (0.002) audio only multiple-talkers
			0.011 (0.002) audio-visual single-talker
			0.008 (0.002) audio-visual multiple-talkers
Modality of presentation	0.011	0.918	0.010 (0.002) audio-only
			0.010 (0.002) audio-visual

Table 2 | Summary of results from the split plot ANOVA [Talker Variability (Single-Talker vs. Multiple-Talkers) × Modality of Presentation (Audio-only vs. Audio-visual), with Talker Variability as a within-subject factor and Modality of Presentation as a between-subject factor] for the dependent measure of hit rates.

Source	*F* statistic	*p*	Estimated means (standard error)
Talker variability	0.199	0.658	0.964 (0.006) single-talker
			0.962 (0.005) multiple-talkers
Talker Variability × Modality of Presentation	0.797	0.377	0.955 (0.008) audio only single talker
			0.957 (0.007) audio only multiple-talkers
			0.973 (0.008) audio-visual single-talker
			0.967 (0.007) audio-visual multiple-talkers
Modality of presentation	1.897	0.176	0.956 (0.007) audio-only
			0.970 (0.007) audio-visual

Table 3 | Summary of results from the split plot ANOVA [Talker Variability (Single-Talker vs. Multiple-Talkers) × Modality of Presentation (Audio-only vs. Audio-visual), with Talker Variability as a within-subject factor and Modality of Presentation as a between-subject factor] for the dependent measure of d-primes.

Source	*F* statistic	*p*	Estimated means (standard error)
Talker variability	0.505	0.481	0 4.351 (0.101) single-talker
			4.289 (0.089) multiple-talker
Talker Variability × Modality of Presentation	0.000	0.988	4.282 (0.143) audio only single-talker
			4.221 (0.125) audio only multiple-talkers
			4.420 (0.143) audio-visual single-talker
			4.357 (0.125) audio-visual multiple-talkers
Modality of presentation	0.653	0.423	4.252 (0.120) audio-only
			4.389 (0.120) audio-visual

1995; Rosenblum et al., 1996; Lachs et al., 2001). Research shows that talker variability hurts recognition accuracy (e.g., Creelman, 1957) and recognition speed (Mullennix and Pisoni, 1990; Magnuson and Nusbaum, 2007) providing what could be viewed as an adverse listening situation. If this impairment of recognition performance is a result of reduced intelligibility due to phonetic uncertainty (cf. Magnuson and Nusbaum, 2007) then converging information about phonetic identity from a speaker's visemes (Skipper et al., 2005) could improve performance. However, the results show that visual information that is coincident with the acoustic information does not lead to faster recognition in a multiple-talker context; rather the presence of a speaker's face appears to increase the talker variability effect. Listeners who additionally saw a talker's face concurrent with hearing a

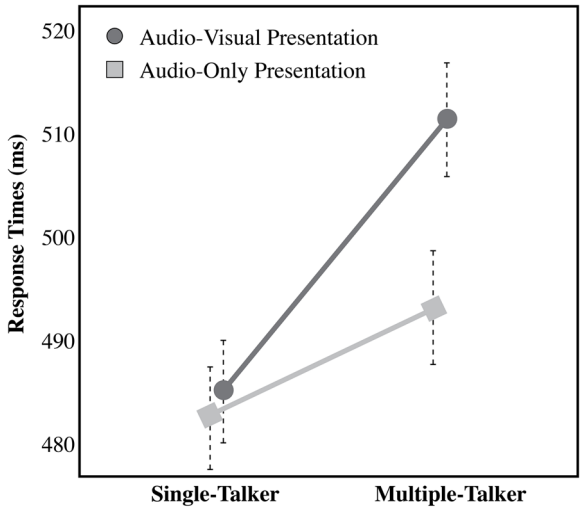


FIGURE 2 | Response times (RTs) for the single-talker and multiple-talker conditions for both presentation modalities (audio-only and audio-visual). Error bars represent 1 SE.

talker were significantly slower to recognize speech in multiple-talker trials compared to single-talker trials and were slowed in this more than listeners who could only heard the speakers. This effect of slowing word recognition for multiple-talker trials when listeners could see each talker however, is not due to the presence of the face alone as there was little difference between audio-only single-talker trials compared audio-visual single-talker trials. For this reason, the exacerbation of the talker variability effect in the audio-visual condition compared to the audio-only condition is not simply a distraction effect of visual information.

The current work only examines the benefits of visual information that is coincident with acoustic information, as all the stimuli across conditions (A and AV) were edited to begin at the start of sound onset. While work by Munhall and Tohkura (1998) demonstrates that visual information is continuously available and incrementally useful to a listener, it is possible that the visual information that precedes the acoustic onset may be helpful in ameliorating the talker variability effect. Work by Smeele (1994, Unpublished Doctoral dissertation) demonstrates that some visual cues can precede the acoustic onset by 80–100 ms. As such, this window may help to prime listeners that a talker change has indeed occurred even before the acoustic signal begins, assuaging the perceptual cost of talker variability. Still, the current work suggests that while visual information that is coincident with acoustic information can influence speech perception (Munhall and Tohkura, 1998), it does not mitigate the short-term accommodation to variability found in a multiple-talker context.

These results are consistent with the perspective that seeing a person speak provides more information about the speaker and the speech than just listening to the speech alone. First, a face conveys clear identifying information, as well as providing information relevant to the message content. Visemes – visual information from mouth shapes (Fisher, 1968) – provide phonetic information, which affects speech perception, and even possess the ability to change what is heard in the acoustic signal as in the McGurk effect. Why does seeing a talker slow recognition even more when there is talker variability? Clearly seeing a talker increases the perception of variability. Even when listeners do not perceive a talker difference in speech (Fenn et al., 2011) seeing the face of a person change in this situation will act as a robust cue that a change in speaker has occurred. When a listener knows that there is a talker change, even when there has been none, there are slowing effects on speech recognition times. Magnuson and Nusbaum (2007) showed that the effect of talker variability is due to the knowledge of a talker change or difference rather than the specifics of an acoustic difference. In the present study, the change in face makes absolutely clear to listeners that there has been a change in talker. In this respect the present results are entirely consistent with previous research.

What is the mechanism by which talker variability interacts with modality? Wong et al. (2004) argued that changes in the talker increased demands on attention in speech processing, showing increased superior parietal activity and increased superior temporal activity. In addition, there was a trend toward increased activity in the premotor system when there was talker variability. Moreover, audio-visual speech perception increases brain activity in the premotor system as well (Skipper et al., 2005). From these results, one could predict that audio-visual talker variability might produce an interaction in activation within perisylvian areas that are involved in speech perception. Such increases in activity might correspond to slower processing rather than faster processing, in that suppression of neural activity by relevant information is usually associated with priming and faster responses (Grill-Spector et al., 2006).

While talker normalization accounts have suggested that slowing due to talker variability is a consequence of using talker vocal characteristics to calibrate phoneme processing in the context of new talker, it has also been suggested that listeners also need to identify talkers for more than just reducing phonetic uncertainty. Labov (1986) has argued that listeners need to understand the social context of a message in order to understand it. For example, Holtgraves (1994) has shown that speech is understood differently depending on the attributed power of the speaker. Rubin (1992) demonstrated that a picture of a putative speaker displaying racial group membership could change the perceived intelligibility of speech. Johnson et al. (1999) have shown that changing expectations about a speaker's gender, just from a static picture of the speaker, can change vowel perception. Niedzielski (1999) has shown that changing listeners' beliefs about a speaker's dialect can change vowel perception. All of these examples reflect the way that knowledge about a speaker's social identity can change speech perception. Although a speaker's social identity can be conveyed through speech by dialect or voice differences, seeing a person's face conveys a great deal more social information. The present results suggest that listeners will process this identifying information even if there is a slight cost in recognition speed,

which may reflect the importance of social information in speech understanding.

ACKNOWLEDGMENTS

The authors would like to thank Chi-Hyun Kim for his assistance in conducting the study.

REFERENCES

Bahrick, H. P., Bahrick, O. O., and Wittlinger, R. P. (1975). Fifty years of memory for names and faces: a cross-sectional approach. *J. Exp. Psychol.* 104, 54–57. doi: 10.1037/0096-3445.104.1.54

Barreda, S., and Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *J. Acoust. Soc. Am.* 131, 466–477. doi: 10.1121/1.3662068

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot Int.* 5, 341–345.

Creelman, C. D. (1957). Case of the unknown talker. *J. Acoust. Soc. Am.* 29, 655. doi: 10.1121/1.1909003

Diamond, R., and Carey, S. (1986). Why faces are and are not special: an effect of expertise. *J. Exp. Psychol.* 115, 107–117. doi: 10.1037/0096-3445.115.2.107

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). Stop-consonant recognition: release bursts and formant transitions as functionally equivalent, context-dependent cues. *Percept. Psychophys.* 22, 109–122. doi: 10.3758/BF03198744

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., and Nusbaum, H. C. (2011). When less is heard than meets the ear: change deafness in a telephone conversation. *Q. J. Exp. Psychol.* 64, 1442–1456. doi: 10.1080/17470218.2011.570353

Fisher, C. G. (1968). Confusions among visually perceived consonants. *J. Speech Hear. Res.* 11, 796–804.

Gauthier, I., and Nelson, C. (2001). The development of face expertise. *Curr. Opin. Neurobiol.* 11, 219–224. doi: 10.1016/S0959-4388(00)00200-2

Gerstman, L. (1968). Classification of self-normalized vowels. Audio and Electroacoustics. *IEEE Trans.* 16, 78–80. doi: 10.1109/TAU.1968.1161953

Goh, W. D., Pisoni, D. B., Kirk, K. J., and Remez, R. E. (2001). Audio-visual perception of sinewave speech in an adult cochlear implant user: a case study. *Ear Hear.* 22, 412–419. doi: 10.1097/00003446-200110000-00005

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295X.105.2.251

Goldinger, S. D., Pisoni, D. B., and Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *J. Exp. Psychol. Learn. Mem. Cogn.* 17:152. doi: 10.1037/0278-7393.17.1.152

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10, 14–23. doi: 10.1016/j.tics.2005.11.006

Halle, M. (1985). "Speculations about the representation of words in memory," in *Phonetic Linguistics*, ed. V. A. Fromkin (New York: Academic Press).

Hasson, U., Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56, 1116–1126. doi: 10.1016/j.neuron.2007.09.037

Holtgraves, T. M. (1994). Communication in context: the effects of speaker status on the comprehension of indirect requests. *J. Exp. Psychol. Lear. Mem. Cogn.* 20, 1205–1218. doi: 10.1037/0278-7393.20.5.1205

Huang, J., and Holt, L. L. (2012). Listening for the norm: adaptive coding in speech categorization. *Front. Psychol.* 3:10. doi: 10.3389/fpsyg.2012.00010

IEEE Subcommittee on Subjective Measurements. (1969). IEEE recommended practices for speech quality measurements. *IEEE Trans. Audio Electroacoust.* 17, 227–246.

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *J. Phon.* 27, 359–384. doi: 10.1006/jpho.1999.0100

Joos, M. (1948). Acoustic phonetics. *Language* 24, 5–136. doi: 10.2307/522229

Labov, W. (1986). "Sources of inherent variation in the speech process," in *Invariance and Variability in Speech Processes*, eds J. Perkell and D. H. Klatt (Hillsdale, NJ: Erlbaum), 402–425.

Lachs, L., Pisoni, D. B., and Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report. *Ear Hear.* 22, 236–251. doi: 10.1097/00003446-200106000-00007

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98–104. doi: 10.1121/1.1908694

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Logan, J. S., Lively, S. E., and Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *J. Acoustic. Soc. Am.* 89, 874–886. doi: 10.1121/1.1894649

Magnuson, J. S., and Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 391–409. doi: 10.1037/0096-1523.33.2.391

Magnuson, J. S., Yamada, R. A., and Nusbaum, H. C. (1994). "Variability in familiar and novel talkers: effects on mora perception and talker identification," in *Proceedings of the Acoustical Society of Japan Technical Committee on Psychological and Physiological Acoustics,* Kanazawa, H-94-44, 1–8.

Massaro, D. W., and Cohen, M. M. (1995). Perceiving talking faces. *J. Acoust. Soc. Am.* 97, 3308–3308. doi: 10.1121/1.412931

McGurk, H., and MacDonald, J. W. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

McLennan, C. T., and Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 306–321. doi: 10.1037/0278-7393.31.2.306

Mullennix, J. W., and Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Percept. Psychophys.* 47, 379–390. doi: 10.3758/BF03210878

Munhall, K. G., and Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *J. Acoust. Soc. Am.* 104, 530–539. doi: 10.1121/1.423300

Munhall, K. G., and Vatikiotis-Bateson, E. (1998). "The moving face during speech communication," in *Hearing by Eye, Part 2: The Psychology of Speech Reading and Audiovisual Speech,* eds R. Campbell, B. Dodd, and D. Burnham (London: Taylor and Francis, Psychology Press), 123–139.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 85, 2088–2113. doi: 10.1121/1.397861

Niedzielski, N. (1999). The effects of social information on the perception of sociolinguistic variables. *J. Lang. Soc. Psychol.* 18, 62–85. doi: 10.1177/0261927X99018001005

Nusbaum, H. C., and Magnuson, J. (1997). "Talker normalization: phonetic constancy as a cognitive process," in *Talker Variability in Speech Processing*, eds K. Johnson and J. W. Mullennix (San Diego, CA: Academic Press), 109–132.

Nusbaum, H. C., and Morin, T. M. (1992). "Paying attention to differences among talkers," in *Speech Perception, Production and Linguistic Structure,* eds Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Tokyo: OHM Publishing Company), 113–134.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychol. Sci.* 5, 42–46. doi: 10.1111/j.1467-9280.1994.tb00612.x

Olsson, N., Juslin, P., and Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *J. Exp. Psychol. Appl.* 4, 101–118. doi: 10.1037/1076-898X.4.2.101

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175–184. doi: 10.1121/1.1917300

Pisoni, D. B. (1993). Long-term memory in speech perception: some new findings on talker variability, speaking rate and perceptual learning. *Speech Commun.* 13, 109–125. doi: 10.1016/0167-6393(93)90063-Q

Pisoni, D. B. (1997). "Some thoughts on "normalization" in speech perception," in *Talker Variability in Speech Processing*, eds K. Johnson and J. W. Mullennix (San Diego, CA: Academic Press), 9–32.

Read, D., and Craik, F. I. M. (1995). Earwitness identification: some influences on voice recognition. *J. Exp. Psychol. Appl.* 1, 6–18. doi: 10.1037/1076-898X.1.1.6

Rosenblum, L. D., Johnson, J. A., and Saldana, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *J. Speech Hear. Res.* 39, 1159–1170.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduate's judgments of nonnative English-speaking teaching assistants. *Res. High. Educ.* 33, 511–531. doi: 10.1007/BF00973770

Schacter, D. L. (1992). Understanding implicit memory: a cognitive neuroscience approach. *Am. Psychol.* 47, 559. doi: 10.1037/0003-066X.47.4.559

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89. doi: 10.1016/j.neuroimage.2004.11.006

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading,* eds B. Dodd and R. Campbell (London: Erlbaum), 3–51.

Summerfield, Q., and Haggard, M. P. (1973). Vocal tract normalization as demonstrated by reaction times. *Rep. Speech Res. Prog.* 2, 12–23.

Syrdal, A. K., and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.* 79, 1086–1100. doi: 10.1121/1.393381

Thakerar, J. N., and Giles, H. (1981). They are – so they spoke : noncontent speech stereotypes. *Lang. Commun.* 1, 255–261. doi: 10.1016/0271-5309(81)90015-X

Wilding, J., and Cook, S. (2000). Sex differences and individual consistency in voice identification. *Percept. Mot. Skills* 91, 535–538. doi: 10.2466/pms.2000.91.2.535

Wong, P. C. M., Nusbaum, H. C., and Small, S. (2004). Neural bases of talker normalization. *J. Cogn. Neurosci.* 16, 1173–1184. doi: 10.1162/08989290419 20522

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., et al. (2009). Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. *Neuroimage* 46, 226–240. doi: 10.1016/j.neuroimage.2009.01.028

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 April 2014; accepted: 17 June 2014; published online: 16 July 2014.
Citation: Heald SLM and Nusbaum HC (2014) Talker variability in audio-visual speech perception. Front. Psychol. ***5****:698. doi: 10.3389/fpsyg.2014.00698*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 08 May 2014
doi: 10.3389/fpsyg.2014.00397

Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders

***Julia R. Irwin*[1,2] * and *Lawrence Brancazio*[1,2]**

[1] *Haskins Laboratories, New Haven, CT, USA*
[2] *Department of Psychology, Southern Connecticut State University, New Haven, CT, USA*

Edited by:
Jean-Luc Schwartz, National Centre for Scientific Research, France

Reviewed by:
Satu Saalasti, Brain and Mind Laboratory, Aalto University School of Science, Finland
David House, Royal Institute of Technology, Sweden

****Correspondence:***
Julia R. Irwin, Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA
e-mail: julia.irwin@haskins.yale.edu

Using eye-tracking methodology, gaze to a speaking face was compared in a group of children with autism spectrum disorders (ASD) and a group with typical development (TD). Patterns of gaze were observed under three conditions: audiovisual (AV) speech in auditory noise, visual only speech and an AV non-face, non-speech control. Children with ASD looked less to the face of the speaker and fixated less on the speakers' mouth than TD controls. No differences in gaze were reported for the non-face, non-speech control task. Since the mouth holds much of the articulatory information available on the face, these findings suggest that children with ASD may have reduced access to critical linguistic information. This reduced access to visible articulatory information could be a contributor to the communication and language problems exhibited by children with ASD.

Keywords: autism spectrum disorders, audiovisual speech perception, eyetracking, communication development, speech in noise, lipreading

INTRODUCTION

Autism spectrum disorders (ASD) refer to neurodevelopmental disorders along a continuum of severity that are generally characterized by marked deficits in social and communicative functioning (American Psychiatric Association, 2000). A feature of the social deficits associated with ASD is facial gaze avoidance and reduced eye contact with others in social situations (Hutt and Ounstead, 1966; Hobson et al., 1988; Volkmar et al., 1989; Volkmar and Mayes, 1990; Phillips et al., 1992). One implication of this reduced gaze to other's faces is a potential difference in face processing. A number of studies have suggested that individuals with ASD show differences in face processing, including impaired face discrimination and recognition (for a review see Dawson et al., 2005, but see Jemel et al., 2006 for evidence that face processing abilities are stronger in ASD than previously reported) and identification of emotion (Pelphrey et al., 2002).

Along with identity and affective information, the face provides valuable information about a talker's articulations. Visible speech information influences what typically developing listeners hear (e.g., increases identification in the presence of auditory noise, Sumby and Pollack, 1954) and is known to facilitate language processing (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978; Reisberg et al., 1987; Desjardins et al., 1997; MacDonald et al., 2000; Lachs and Pisoni, 2004). Further, typical speech and language development is thought to take place in an audiovisual (AV) context (Meltzoff and Kuhl, 1994; Desjardins et al., 1997; Lachs et al., 2001; Bergeson and Pisoni, 2004). Thus, differences in access to visible speech information would have significant consequences for a perceiver. For example, there is evidence that the production of speech differs in blind versus sighted individuals (for example, sighted speakers produce vowels further apart in articulatory space than those of blind speakers, ostensibly because of their access to visible contrasts; Menard et al., 2009), suggesting that speech perception and production is influenced by experience with the speaking face.

Consistent with their difficulties with information on faces, a growing body of literature indicates that children with ASD are less influenced by visible speech information than TD controls (De Gelder et al., 1991; Massaro and Bosseler, 2003; Williams et al., 2004; Mongillo et al., 2008; Iarocci et al., 2010; Irwin et al., 2011, but see Iarocci and McDonald, 2006 and Woynaroski et al., 2013). In particular, children and adolescents with ASD appear to benefit less from the visible articulatory information on the speaker's face in the context of auditory noise (Smith and Bennetto, 2007; Irwin et al., 2011). Further, children with ASD have been reported to be particularly poor at lipreading (Massaro and Bosseler, 2003).

Although avoidance of gaze to others' faces has been noted clinically, the exact nature of gaze patterns to faces in ASD has been a topic of investigation. A varied body of research using eye-tracking methodology has examined patterns of facial gaze patterns in individuals with ASD, in particular with complex social situations and with affective stimuli. A number of studies find that individuals with ASD differ in the amount of fixations to the eye region of the face when compared to typically developing (TD) controls (Klin et al., 2002; Pelphrey et al., 2002; Dalton et al., 2005; Boraston and Blakemore, 2007; Speer et al., 2007; Kleinhans et al., 2008; Sterling et al., 2008). In particular, during affective or emotion based tasks, individuals with ASD have been reported to spend significantly more time looking at the mouth (Klin et al., 2002; Neumann et al., 2006; Spezio et al., 2007). However, a recent review by Falck-Ytter and von Hofsten (2011) calls into question whether individuals with ASD look less to the eyes and more to the mouth when gazing at faces; they argue that only limited support exists for this in adults and even less evidence in children. Apart from gaze to eyes and mouth, some studies show increased gaze at "non-core" features (e.g., regions other than the eyes, nose, and mouth) of the face by

individuals with ASD compared to TD controls, when gazing at facial expression of emotion (Pelphrey et al., 2002). Reports of differences in patterns of gaze to faces are not unequivocal, however, with a number of studies reporting no group differences in certain tasks (Adolphs et al., 2001; Speer et al., 2007; Kleinhans et al., 2008). Further, when assessing gaze to a face, pattern of gaze may be a function of both language skill and development. Norbury et al. (2009) report that pattern of gaze to the mouth is associated with communicative competence in ASD. Reported differences in gaze to faces in children with ASD appear to vary depending on the age of the child (Dawson et al., 2005; Chawarska and Shic, 2009; Senju and Johnson, 2009). Moreover, recent work by Foxe et al. (2013) suggests that multisensory integration deficits present in children with ASD may resolve in adulthood (although subtle differences may persist; Saalasti et al., 2012).

Critically, little is known about gaze to the face during speech perception tasks. A question that arises is whether the previously reported deficit in visual speech processing in children with ASD might simply be a consequence of a failure to fixate on the face. However, recent findings by Irwin et al. (2011) provide evidence against this possibility. Irwin et al. (2011) tested children with ASD and matched TD peers on a set of AV speech perception tasks while concurrently recording eye fixation patterns. The tasks included a speech-in-noise task with auditory-only (static face) and AV syllables (to measure the improvement in perceptual identification with the addition of visual information), a McGurk task (with mismatched auditory and visual stimuli), and a visual-only (speechreading) task. Crucially, Irwin et al. (2011) excluded all trials where the participant did not fixate on the speaker's face. They found that even when fixated on the speaker's face, children with ASD were less influenced by visible articulatory information than their TD peers, both in the speech-in-noise tasks and with AV mismatched (McGurk) stimuli. Moreover, the children with ASD were less accurate at identifying visual-only syllables than the TD peers (although their overall speechreading accuracy was fairly high).

Irwin et al.'s (2011) findings indicate that fixation on the face is not sufficient to support efficient AV speech perception. This could suggest differences in how visual speech information is processed in individuals with ASD. However, it could also be due to different gaze patterns on a face exhibited by individuals with ASD. Perhaps if they tend to fixate on different regions of the face than TD individuals, individuals with ASD have reduced access to critical visual information. Consistent with this possibility is evidence that attentional factors can modulate visual influences in speech perception in typical adults; visual influence is reduced when perceivers are asked to attend to a distractor stimulus on the speaker's face (Alsius et al., 2005). Typically developing adults have been shown to increase gaze to the mouth area of the speaker as intelligibility decreases during AV speech tasks (Yi et al., 2013). Further, Buchan et al. (2007) report that typically developing adults gaze to a central area on the face in the presence of AV speech in noise, reducing the frequency of gaze fixations on the eyes and increasing gaze fixations to the nose and the mouth. If children with ASD do not have access to the same visible articulatory information as the TD controls because their gaze patterns differ, this may influence their perception of a speaker's message.

To assess whether there are differences in gaze that underlie the AV speech perception differences in children with ASD as compared to children with typical development, for the present paper we conducted a detailed analysis of the eye-gaze patterns for the participants and tasks reported in Irwin et al. (2011). In particular, we examined patterns of gaze to a speaking face under perceptual conditions where there is an incentive to look at the face: (1) in the presence of auditory noise and (2) where no auditory signal is present (speechreading). We tested whether children with ASD differ from TD controls not only in overall time spent on the face, but also in the relative amount of time spent fixating on the mouth and non-focal regions. We further examined whether the two groups differ in the time-course of eye-gaze patterns to these regions over the course of a speech syllable. Given that the children with ASD in this sample exhibited poorer use of visual speech information than the TD controls in perceptual measures (both for visual-only and AV speech), the analyses reported here may shed some light on the basis for these differences: Is reduced use of visual speech information in perception associated with differences in patterns of fixation on the talking face?

Finally, as a control for the possibility that there are more general group differences in gaze pattern unrelated to faces, we also analyzed gaze patterns in a control condition with dynamic AV non-face, non-speech stimuli.

MATERIALS AND METHODS

PARTICIPANTS

Participants in the current study were 20 native English speaking monolingual children, 10 with ASD (eight boys, mean age 10.2 years, age range 5.58–15.9 years) and 10 TD controls (eight boys, mean age 9.6, age range 7–12.6 years). Because the speech conditions in this study required the child participants to report what the speaker said, all participants in this study were verbal. All child participants were reported by parents to have normal or corrected-to-normal hearing and vision. The TD participants had no history of developmental delays including vision, hearing, speech or language problems, by parent report.

The TD controls were matched with the child ASD participants on sex, age, cognitive functioning and language skill. The TD controls were taken from a larger set of children participating in a study of speech perception ($n = 80$). In addition, the primary caregivers of children with ASD completed a diagnostic interview [autism diagnostic interview-revised (ADI-R), Lord et al., 1994] about their children ($n = 10$ adult females).

Prior to their participation in the study, child participants with ASD received a diagnosis from a licensed clinician. Four participants had a diagnosis of autism, four of Asperger syndrome and two were diagnosed with pervasive developmental disorder not otherwise specified (PDD-NOS); these diagnoses all fall within the classification of ASD. For characterization purposes, participants with ASD were also assessed with the autism diagnostic observation schedule (ADOS; Lord et al., 2000), and their caregivers ($n = 10$) were interviewed with the ADI-R (Lord

et al., 1994). All participants with ASD met or exceeded cutoff scores for autism spectrum or autism proper on the ADOS algorithm. Scores obtained from caregiver interviews showed that the children with ASD met or exceeded cutoff criteria on the language/communication, reciprocal social interactions and repetitive behavior/interest domains on the ADI-R. Consistent with the range of clinical diagnoses, there was heterogeneity in the extent of social and communication deficits and presence of restricted and repetitive behavior (for example, scores on the combined communication and social impairment scales in the ADOS ranged from 7 to 20, where 10 is the minimum cutoff score and 22 is the maximum possible score).

The mean age and standard deviations of the child ASD and child TD participants, along with measures of cognitive and language functioning, are presented in **Table 1**. The measures of cognitive functioning were standardized scores for general conceptual ability (GCA) on the Differential Abilities Scale (DAS); the measures of language function were core language index scores (CLI) from the clinical evaluation of language fundamentals-4 (CELF-4; Semel et al., 2003). Independent-samples *t*-tests on age, GCA, and CLI did not reveal significant differences between the groups, as shown in **Table 1**.

The sample included here represents a subset of the participants whose data were reported in Irwin et al. (2011). The data of three children with ASD and one TD control were excluded from the present analyses because they spent too little time fixating on the face to permit statistical analysis. The data of two other TD control participants were also removed due to the removal of their respective matched ASD participants.

MATERIALS

Stimuli

Speech stimuli. The speech stimuli were created from a recording of the productions of a male, monolingual, native speaker of American English. This speaker was audio- and video-recorded in a recording booth producing a randomized list of the consonant-vowel (CV) syllables /ma/ and /na/. The video was centered on the speaker's face and was framed from just above the top of the speaker's head to just below his chin, and was captured at 640 × 480 pixels. The audio was simultaneously recorded to computer and normalized for amplitude, and then realigned with the video in Final Cut Pro. Two tokens of /ma/ and /na/ were selected as stimuli. The stimuli were trimmed to start with the mouth position at rest, followed by an opening gesture, closing for the consonant, and release of the consonant into the following vowel, and ended with the mouth returning to rest at the end of the syllable. The stimuli were approximately 1500 ms long, with the acoustic onset of the consonant (for the AV stimuli) occurring at around 600 ms; the acoustic portions of the stimuli were approximately 550 ms in duration, on average.

For AV speech in noise, the stimuli were AV stimuli of /ma/ and /na/. Three versions of each stimulus was created by setting the mean dB of the syllables at 60 dBA, and then adding pink noise at 70, 75, and 70 dBA to the AV /ma/ and /na/ tokens to create stimuli with a range of signal-to-noise levels from less to more noisy (i.e., −10, −15, and −20 dB S/N, respectively). Noise onset and offset were aligned to the auditory speech syllable onset and offset.

The visual-only (speechreading) stimuli were identical to the AV stimuli, except that the audio channel was removed.

Non-speech control stimuli. The AV non-speech stimuli consisted of a set of figure-eight shapes that increased and decreased in size, paired with sine-wave tones that varied in frequency and amplitude. These stimuli were modeled on the speaker's productions of /ma/ and /na/ but did not look or sound like speech. To create the visual stimulus, we measured the lip aperture in every video frame of the /ma/ and /na/ syllables. We then used the aperture values to drive the size of the figure: when the lips closed the figure was small, and upon consonant release into the vowel the figure expanded (see **Figures 1C,D**). The auditory stimuli were created by converting the auditory /ma/ and /na/ syllables into sine-wave analogs, which consist of three or four time-varying sinusoids, following the center-frequency and amplitude pattern of the spectral peaks of an utterance (Remez et al., 1981). These sine-wave analogs sound like chirps or tones. Thus, the AV non-speech stimuli retained the temporal dynamics of speech, without looking or sounding like a speaking face (see **Figures 1A–E**).

Table 1 | Mean age and cognitive and language measures for the children with ASD and TD.

	ASD	TD	*T*-test
n	10	10	
Age	10.2 (3.1)	9.6 (2.4)	$t(18) = -0.51$, ns
General conceptual ability (GCA)	92.1 (15.5)	98.9 (15.5)	$t(18) = 0.97$, ns
Core language index scores (CLI)	87.4 (17.3)	97.8 (15.1)	$t(18) = 1.4$, ns

GCA and CLI are standardized scores. Standard deviations are in parentheses.

Visual tracking methodology

Visual tracking was done with an ASL Model 504 pan/tilt remote tracking system, a remote video-based single eye tracker that uses bright pupil, coaxial illumination to track both pupil and corneal reflections at 120 Hz. To optimize the accuracy of the pupil coordinates obtained by the optical camera, this model has a magnetic head tracking unit that tracks the position of a small magnetic sensor attached to the head of the participant, above their left eye.

Language assessment

Language ability was assessed with the CELF-4 (Semel et al., 2003). The CELF-4 is reliable in assessing the language skills of children in the general population and those with a clinical diagnosis including ASD (Semel et al., 2003).

Cognitive assessment

Cognitive ability was assessed using the Differential Ability Scales (DAS) School Age Cognitive Battery (Elliott, 1991). The DAS provides a GCA score, which assesses verbal ability, non-verbal reasoning ability, and spatial ability.

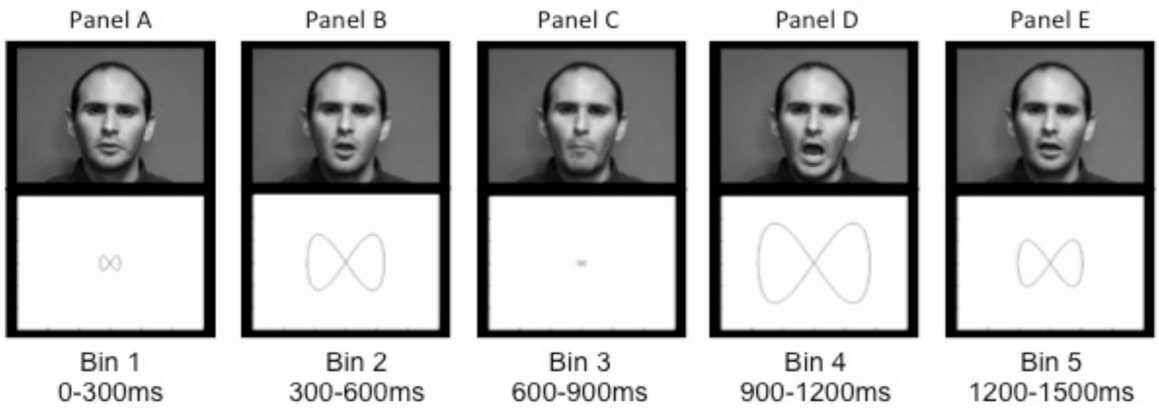


FIGURE 1 | Sample images of the speaker (top panels) during a production of /ma/ and the corresponding non-speech figure-eight shapes (lower panels) taken from each time bin. Panels A through E illustrate, respectively, the initial rest position **(A)**, opening prior to the consonant closing gesture **(B)**, the closure for /m/ **(C)**, peak mouth opening for the vowel **(D)**, and the return to rest at the end of the vowel **(E)**.

ADOS

Children with ASD were assessed with the ADOS generic (ADOS-G). The ADOS is a semi-structured standardized assessment of communication, social interaction, and play/imaginative use of materials for individuals suspected of having an ASD (Lord et al., 2002).

ADI-R

Caregivers of participants with ASD were given the ADI-R (Lord et al., 1994). The ADI-R is a standardized, semi-structured interview for caregivers of those with an ASD to assess autism symptomatology.

PROCEDURE

After consent was obtained in accordance with the Yale University School of Medicine, all participants completed the experimental tasks in the eye-tracker. Each participant was placed in front of the monitor, after which calibration of the participant's fixation points in the eye-tracker was completed. Prior to any stimulus presentation for each task, directions appeared on the monitor. These directions were read aloud to the participant by a researcher to ensure that they understood the task. In addition, two practice items were completed with the researcher present to confirm that the participant understood and could complete the task. For all conditions, if participants were unsure, they were asked to guess.

Condition 1: AV speech in noise

Participants were told that they would see and hear a man saying some sounds that were not words and to say out loud what they heard. Each of the six stimuli (two different tokens of each /ma/ and /na/, at each of the three levels of signal-to-noise ratios) was presented four times, for a total of 24 trials in a random sequence.

Condition 2: visual only (speechreading)

Participants were told that they would see a man saying some sounds that they would not be able to hear, and then asked to say out loud what they thought the man was saying. Each of the four stimuli (two different tokens of each /ma/ and /na) was presented five times, for a total of 20 trials in a random sequence.

Condition 3: non-speech control

For this task, two stimuli were presented in sequence on each trial. The paired stimuli were either modeled on different tokens of the same syllable (e.g., both /ma/ or both /na/) or on tokens of different syllables (one /ma/ and one /na/). Participants were told that they would see two shapes that would open and close and should say out loud whether the two shapes opened and closed in the same way (e.g., both modeled on /ma/ or both modeled on /na/, although no reference was made to the speech origins of the stimuli to participants) or if the way that they closed was different (e.g., one modeled on /ma/ and one on /na/). Each pairing was presented seven times, for a total of 28 trials in a random sequence.

The three tasks were blocked and presented in random order. The inter-stimulus interval for all trials within the blocks was 3 s. After every five trials, participants were presented with a slide of animated shapes and faces, to maintain attention to the task. All audio stimuli were presented at a comfortable listening level (60 dBA) from a centrally located speaker under the eye-tracker, and visual stimuli were presented at a 640 × 480 aspect ratio on a video monitor 30 inches from the participant.

After the experimental procedure participants were tested with the battery of cognitive and language assessments and caregivers of the ASD participants were interviewed separately with the ADI-R.

RESULTS

Participant gaze to the speaker's face was examined by group for the AV speech-in-noise and visual-only (speechreading) trials, as was gaze on the figure-eight shape in non-speech trials. The eye tracker recorded fixation position in x and y coordinates at approximately 8 ms intervals. (In cases where the coordinates were not recorded, the x- and y-coordinates of the previous time point were applied).

Each x-y coordinate was coded according to whether it was on-screen or off-screen, and if it was on-screen, whether it was part of an on-face fixation or not. Off-screen fixations were eliminated from the data.

The on-face coordinates were coded according to face regions, namely: forehead, jaw, cheeks, ears, eyes, mouth region (including the spaces between the lower lip and the jaw and between the upper lip and the nose), and nose. The primary regions of interest were the *mouth region* and a collective set of *non-focal regions* (face areas other than the mouth region, eyes, and nose), in light of reports that children with ASD spend relatively more time fixating on non-focal regions of the face (Pelphrey et al., 2002). The non-focal regions encompassed the ears, the cheeks, the forehead, and all other regions not otherwise labeled (primarily the space between the eye and the ear, between the nose and cheek, and between the eyes). The jaw area was not included in either the mouth region or the non-focal regions; this is because the jaw, unlike the other non-focal regions, has extensive movement that is time-locked to the speech articulation – thus, jaw movement conveys information about the kinematics of the speech act.

For the non-speech condition, the on-screen regions were coded in an analogous manner, based on the extent of the figure-eight shape. These regions are described below.

Data points were only included as fixations if they had less than a 40 pixel movement from the previous time point, and occurred within a contiguous 100 ms window of similar small movements that did not cross into a different face region, as defined above. In all, 14.5% of the time steps were eliminated across the AV speech-in-noise and visual-only tasks for being either off-screen, saccades, or blinks. Although the mean percentage of dropped data points was higher for the ASD sample than for the TD sample, the difference was not statistically significant [for AV speech-in-noise, ASD: $M = 19.4\%$, SD $= 13.3$; TD: $M = 11.8\%$, SD $= 7.4$; $t(18) = 1.60$, ns; for visual-only, ASD: $M = 17.0\%$, SD $= 12.0$; TD: $M = 10.0\%$, SD $= 5.3$; $t(18) = 1.70$, ns].

The individual time steps were collapsed into 300 ms time bins (0–300 ms, 300–600 ms, 600–900 ms, 900–1200 ms, and 1200–1500 ms); we thus calculated the total amount of time spent in each region within each time bin. These time bin boundaries were selected because they roughly corresponded to visual landmarks in the speech signal. The first bin (0–300 ms) preceded the onset of visible movement; the second bin (300–600 ms) included opening of the mouth prior to the consonant and the initiation of closing (either lips in /ma/ or upward tongue-tip movement in /na/); the third bin (600–900 ms) included the consonantal closure and release, and the final two time bins (900–1200 ms and 1200–1500 ms, respectively) span production of the vowel until the end of the trial (for an image of articulation in each of the time bins paired with the corresponding figure-eight shape, see **Figure 1**).

As a result, our dependent variables were the mean percentage of time gazing on a given region within a time bin. Time spent fixating on the *face* was calculated as a percentage of time fixated anywhere on the computer monitor within each time bin. In contrast, time spent fixating on *specific face regions* (mouth region and non-focal areas) was calculated as a percentage of time spent fixated on the face within each time bin.

First, we examined whether there were group differences in the percentage of time spent fixating on the *face* of the speaker out of time spent fixating on-screen. **Figure 2** presents the mean time spent on face by group and time bin separately for the AV speech-in-noise and visual-only tasks. As the figure shows, the ASD group on average spent consistently less time on the face than the TD group in both tasks. A set of 2 (group: ASD, TD) by 5 (time bin: 0–300 ms, 300–600 ms, 600–900 ms, 900–1200 ms, and 1200–1500 ms) mixed factor analyses of variance (ANOVAs) were conducted for AV speech-in-noise and visual-only, respectively. There was a significant main effect of group with less time spent on the face by the ASD group than the TD group for AV speech-in-noise with a marginal effect for visual-only [for AV speech-in-noise, ASD: $M = 60.8$, SD $= 25.0$; TD: $M = 82.3$, SD $= 21.9$; $F(1,18) = 6.31$, $p = 0.02$, $\eta^2_G = 0.22$; for visual-only, ASD: $M = 74.3$, SD $= 20.7$; TD: $M = 84.2$, SD $= 14.9$; $F(1,18) = 3.39$, $p = 0.08$, $\eta^2_G = 0.12$]. These mean differences reflect moderate to large effect size estimates (Cohen, 1973; Olejnik and Algina, 2003; Bakeman, 2005). There was also a main effect of time bin in both analyses [AV speech-in-noise: $F(4,72) = 26.48$, $p < 0.0001$, $\eta^2_G = 0.23$; visual-only: $F(4,72) = 42.7$, $p < 0.001$, $\eta^2_G = 0.41$], reflecting a rapid increase in fixations on the face from the first to second bins that leveled off by the third bin. The interaction of group and time was not significant for either task.

Next, we examined whether there were group differences in gaze to specific regions on the face. We chose the mouth region and non-focal areas (as defined above) as regions of interest[1]. We ran a set of 2 (group: ASD, TD) by 5 (time bin: 0–300 ms, 300–600 ms, 600–1200 ms, 1200–1500 ms) ANOVAs on the percentage of time spent in each region of interest out of time spent on the face, with separate analyses for the AV speech-in-noise and visual-only tasks, and separate analyses for the mouth region and non-focal areas. **Figure 3** presents the relative percentages of time spent in each region of interest by group and time, separately for the AV speech-in-noise and visual-only tasks.

First, consider the *mouth region*. There was a significant main effect of group for both tasks, with a relatively smaller percentage of time spent on the mouth region for the ASD group than the TD group [for AV speech-in-noise, ASD: $M = 26.0$, SD $= 24.1$; TD: $M = 52.9$, SD $= 30.8$; $F(1,18) = 11.25$, $p < 0.005$, $\eta^2_G = 0.29$; for visual-only, ASD: $M = 35.0$, SD $= 29.5$; TD: $M = 56.1$, SD $= 32.6$; $F(1,18) = 4.46$, $p = 0.05$, $\eta^2_G = 0.14$]. There was also a main effect of time for both tasks [AV speech-in-noise: $F(4,72) = 23.18$, $p < 0.0001$, $\eta^2_G = 0.32$; visual-only: $F(4,72) = 23.7$, $p < 0.0001$, $\eta^2_G = 0.30$], with an overall increase in fixations on the mouth region from the first to third bins before leveling off. Interestingly, there was an interaction of group and time bin for AV speech-in-noise [$F(4,72) = 10.06$, $p < 0.0001$, $\eta^2_G = 0.17$], but not for visual-only ($F < 1$). As shown in **Figure 3**, for AV speech-in-noise, fixations on the mouth region were similar for the two groups in the first time bin (0–300 ms, prior to the onset of mouth movement), but the subsequent increase in mouth region fixations was

[1]In addition to the analyses of the mouth region and non-focal regions, we also conducted statistical analyses of fixations on other major face areas, namely the eyes and nose. However, each involved few fixations overall and the analyses did not reveal reliable differences between groups; thus, they are not reported here.

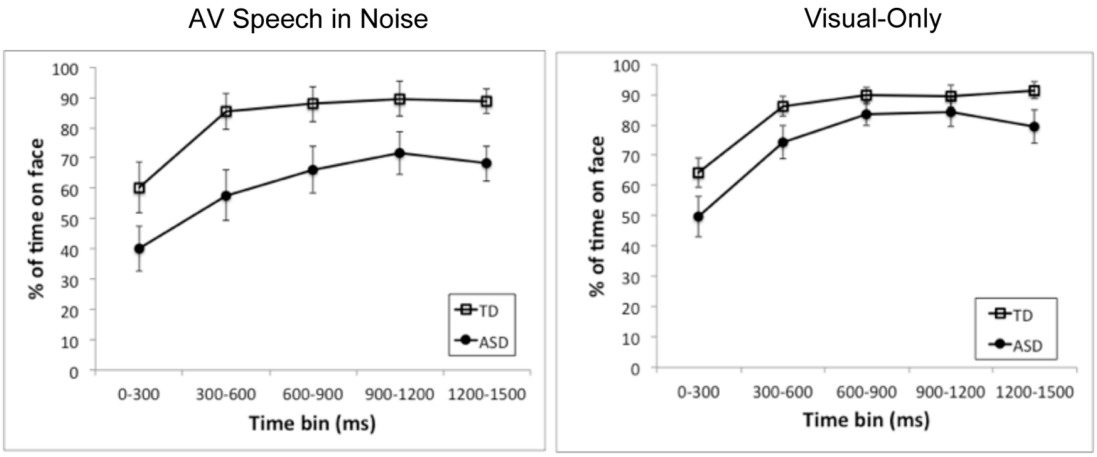


FIGURE 2 | Mean time spent on the face region as a percentage of time spent on-screen for each of the time bins and for the ASD group (closed circles) and the TD group (open squares). The left and right panels present results for AV speech in noise and visual-only, respectively. Error bars represent standard errors, calculated independently for each time bin.

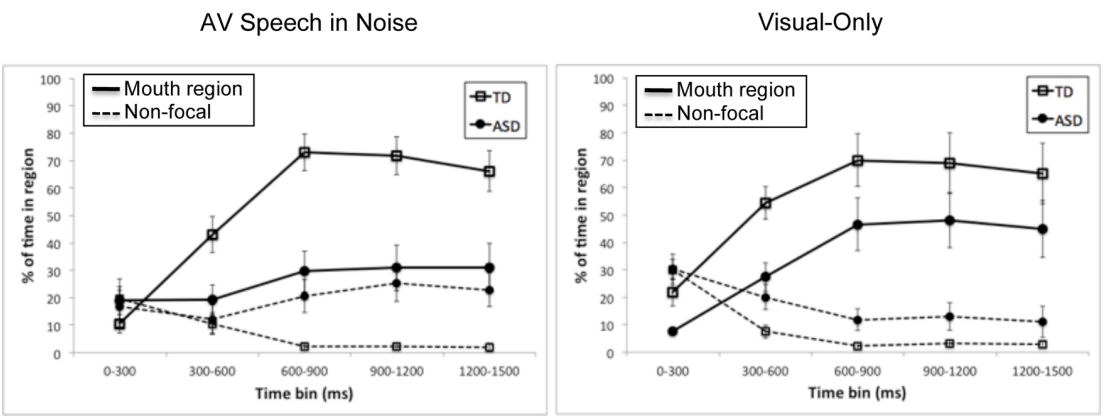


FIGURE 3 | Mean time spent on the mouth region (solid lines) and non-focal areas (dashed lines) as a percentage of time spent on the face for each of the time bins and for the ASD group (closed circles) and the TD group (open squares). The left and right panels present results for AV speech in noise and visual-only, respectively. Error bars represent standard errors, calculated independently for each time bin.

much more pronounced for the TD group than the ASD group. In contrast, in the visual-only task the two groups' trajectories across time were similar, differing in overall percentage of time in the mouth region.

Next, consider the *non-focal regions*. For AV speech-in-noise, there was a significant main effect of group, with a relatively higher percentage of time spent fixating on non-focal regions by the ASD group than the TD group [ASD: $M = 19.5$, SD $= 19.6$; TD: $M = 7.3$, SD $= 10.5$; $F(1,18) = 6.48$, $p < 0.05$, $\eta_G^2 = 0.15$]. There was not a significant main effect of time, $F(4,72) = 1.11$, ns, but there was a significant interaction of group and time, $F(4,72) = 4.98$, $p < 0.005$, $\eta_G^2 = 0.12$. Time spent on non-focal regions was similar for the two groups in the first time bin, but dropped off rapidly for the TD group while remaining relatively frequent for the ASD group across the whole trial. For visual-only, there was again a main effect of group [ASD: $M = 17.3$, SD $= 16.9$; TD: $M = 9.2$, SD $= 12.6$; $F(1,18) = 5.43$, $p < 0.05$, $\eta_G^2 = 0.11$], along with a significant main effect of time, $F(4,72) = 17.64$, $p < 0.0001$, $\eta_G^2 = 0.37$, with a decrease in time spent on non-focal regions from the first time bin to the subsequent bins. The interaction of group and time ($F < 1$) was not statistically significant in the visual-only task[2].

[2]We initially considered the jaw as a non-focal region, but removed it from the category because of its extensive movement during the speech event (thus providing information about the kinematics of the speech act), which distinguished it from other non-focal areas. However, we did repeat the analyses of the non-focal regions with the jaw included. This inclusion did not change the outcome for AV speech-in-noise, but it did for visual-only. In the visual-only task, there were considerably more fixations in the jaw region by the TD participants than the ASD participants (although, in an analysis of just fixations on the jaw, the difference was not statistically reliable). As a result, including jaw in the non-focal category had the effect of eliminating the statistically significant group difference in non-focal fixations. However, this obscures an interesting difference between the groups: The ASD group spent relatively more time fixating on face areas that convey less information about the kinematics of the speech articulations (e.g., the cheeks).

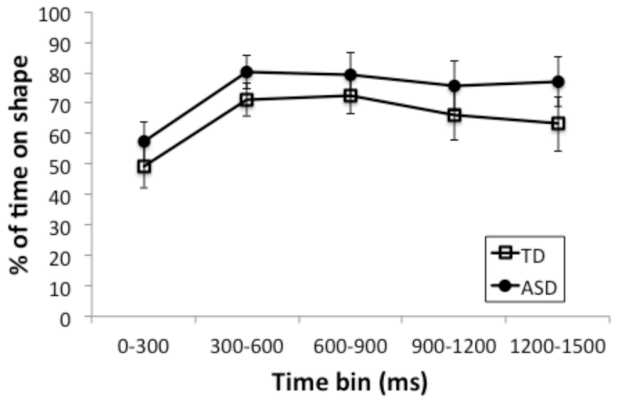


FIGURE 4 | Mean time spent on the figure-eight shape region as a percentage of time spent on-screen for each of the time bins and for the ASD group (closed circles) and the TD group (open squares). Error bars represent standard errors, calculated independently for each time bin.

The results in the speech tasks can be summarized as follows. First, the ASD group spent, on average, less time gazing on the face than the TD group, and this difference was more pronounced in the AV speech-in-noise task than in the visual-only task. Second, when fixating on the face, the ASD group spent relatively less time fixating on the mouth region than the TD group, and relatively more time fixating on non-focal regions. Finally, the two groups differed in their relative pattern of fixations on the speech over the course of a trial. Specifically, the TD group exhibited a pattern of initially looking at non-focal regions but then shifting to the mouth as the articulation unfolded. The ASD group had a similar but reduced shift in the visual-only task, but did not exhibit this shift in the AV speech-in-noise task.

NON-SPEECH CONTROL CONDITIONS

Finally, to assess whether there were group differences in gaze to the non-speech stimuli, a series of independent 2 (group: ASD, TD) × 5 (time bins: 0–300 ms, 300–600 ms, 600–900 ms, 900–1200 ms, and 1200–1500 ms) ANOVAs were run on fixations to the figure-eight shapes during time spent on screen. The earliest time bin encompasses pre-movement (0–300 ms), the next time bin (300–600 ms) an increase to maximum size; the third time bin (600–900 ms) from maximum size to minimum size and the final two time bins increasing until the end of the trial (900–1200 ms, 1200–1500 ms, see **Figure 1**). We defined two regions of interest: a narrow region encompassing an area around the outline of the figure-eight shape at its smallest point (see **Figure 1C**), and a broad region encompassing the area around the outline of the shape at its largest point (see **Figure 1D**). We analyzed percentage of trials with fixations in each region at the previously defined time samples that incorporated the shape's transition from a small outline to a large one. The percentage of time spent in the broad region, shown in **Figure 4**, had a main effect of time bin [$F(4,72) = 12.33$, $p < 0.0001$, $\eta^2_G = 0.13$], due to an increase from the first bin (prior to movement) to the second, but no main effect of group [$F(1,18) = 1.09$, ns] and no interaction of group and time bin ($F < 1$). The percentage of time in the narrow region also had a main effect of time bin [$F(4,72) = 8.32$, $p < 0.001$, $\eta^2_G = 0.14$], with less time in the inner region in the first bin (prior to movement) and in the last two bins (when the shape was larger), but again with no main effect of group ($F < 1$) and no interaction of group and time bin [$F(4,72) = 1.10$, ns]. Overall, the TD and ASD groups exhibited similar gaze patterns with the non-speech stimuli.

DISCUSSION

The current study examined pattern of gaze to a speaking face by children with ASD and a set of well-matched TD controls. Gaze was examined under conditions that create a strong incentive to attend to the speaker's articulations, namely, AV speech with background noise and visual only (speechread) speech. We found differences in the gaze patterns of children with ASD relative to their TD peers, which could impact their ability to obtain visible articulatory information.

The findings indicated that children with ASD spent significantly less time gazing to a speaking face than the TD controls, which is consistent with diagnostic criteria for this disorder and findings from previous research (Hutt and Ounstead, 1966; Hobson et al., 1988; Volkmar et al., 1989; Volkmar and Mayes, 1990; Phillips et al., 1992). The reduction in gaze to the face of the speaker was greater in the speech in noise than the visual-only condition. This suggests that children with ASD gaze at the face of the speaker when the task requires it, as in speechreading. This is perhaps consistent with the finding that the difference in perceptual performance between the ASD and TD groups (Irwin et al., 2011) was less pronounced in the visual-only condition than with speech in noise.

Importantly, when fixated on the face of speaker, the children with ASD were significantly less likely to gaze at the speaker's mouth than the TD children in the context of both speech in noise and speechreading. This finding might appear to conflict with previous findings of increased gaze to the mouth by individuals with ASD in comparison to TD controls (e.g., Klin et al., 2002; Neumann et al., 2006; Spezio et al., 2007). However, this disparity may arise from the specific demands of the respective tasks. Findings of increased gaze on the mouth by children with ASD have typically occurred when the task required emotional or social judgments and when the mouth was not the primary source of the relevant information. In contrast, our study involved a speech perception task, so the mouth *was* the primary source of relevant (articulatory) information. These findings in tandem suggest that children with ASD paradoxically may be *less* likely to attend to the mouth when it carries *greater* informational value.

Instead of gazing at the mouth during the speech in noise task, the children with ASD tended to spend more time directing their gaze to non-focal areas of the face (also see Pelphrey et al., 2002). Non-focal areas such as the ears, cheeks, and forehead carry little, if any, articulatory information. For speech in noise, as the speaker began to produce the articulatory signal, the TD children looked more to the mouth than did the children with ASD, who continued to gaze at non-focal regions.

Notably, the group differences were less prominent in the visual-only condition, where visual phonetic information on the

mouth is fundamental to the task (in contrast to the speech-in-noise task, where there is an auditory speech signal). In this case, the two groups exhibited a similar pattern of shifting from non-focal areas to the mouth region as the speaker began to produce the syllable, even though the ASD group overall spent relatively less time on the mouth and more time on non-focal regions than the TD controls. This finding suggests that children with ASD may be able to approximate a similar pattern of gaze to areas of the face that hold important articulatory information when it is required by the task.

Finally, there were no significant differences by group in pattern of gaze for the non-speech, non-face control condition. This suggests that the differences in gaze patterns between children with ASD and TD do not necessarily occur for all AV stimuli, and are consistent with the notion that these differences are specific to speaking faces.

In the Introduction, we outlined two possible reasons for why children with ASD are less influenced by visual speech information than their TD peers, even when they are fixated on the face (Irwin et al., 2011), namely, that they have an impairment in AV speech processing, or that they have reduced access to critical visual information. The present results do not address the question of a processing impairment, but they do offer insight into the issue of access to speech information. Because the mouth is the source of phonetically relevant articulatory information available on the face (Thomas and Jordan, 2004), our results may help account for the language and communication difficulties exhibited by children with ASD.

To summarize, even with a sample of verbal children who were closely matched in language and cognition to controls, we found differences in pattern of gaze to a speaking face between children with ASD and TD controls. However, these findings should be interpreted with caution, given the small sample size, broad age range and varied diagnostic category. Future research should be conducted to assess how differences in each of these variables impacts pattern of gaze. In particular, an interesting question is whether pattern of gaze relates to communicative skill (e.g., as in Norbury et al., 2009; also see Falck-Ytter et al., 2012). A larger sample would allow for examination of this relationship. Further, the speech stimuli in the current study were consonant-vowel speech syllables; future research should also examine sentence level connected speech.

Finally, future work should consider the possible implications of the results for intervention. Our results in the speech-in-noise task indicate that children with ASD may not spontaneously look to critical areas of a speaking face in the presence of background noise, even though it would improve comprehension. This is particularly problematic in light of findings that auditory noise is especially disruptive for individuals with ASD in speech perception (Alcántara et al., 2004). However, the results in the visual-only speechreading task, where children with ASD did tend to shift their gaze from non-focal areas to the mouth (albeit to a lesser degree than the TD controls), suggests that children with ASD can show more typical gaze patterns when necessary. Therefore, intervention to help individuals with ASD to gain greater access to visible articulatory information may be useful, with the goal of increased communicative functioning in the natural listening and speaking environment.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health. [P01 HD-001994, Jay Rueckl, PI; R03 DC-007339, Julia Irwin, PI; R21 DC-00403, DC-011342, Julia Irwin, PI; R01 DC-000403, Cathi Best, PI]. Thanks go to Jim Magnuson and Alice Carter for discussions about the manuscript. Additional thanks go to Lauren Tornatore for assistance with data collection, and Jessica Ross, Stephanie Murana and Dana Albert for assistance in preparing the manuscript.

REFERENCES

Adolphs, R., Sears, L., and Piven, J. (2001). Abnormal processing of social information from faces in autism. *J. Cogn. Neurosci.* 13, 232–240. doi: 10.1162/089892901564289

Alcántara, J. I., Weisblatt, E. J. L., Moore, B. C. J., and Bolton, P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *J. Child Psychol. Psychiatry* 45, 1107–1114. doi: 10.1111/j.1469-7610.2004.t01-1-00303.x

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.*15, 839–843. doi: 10.1016/j.cub.2005.03.046

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, DSM-IV-TR®*. Washington, DC: American Psychiatric Publishing.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behav. Res. Methods* 37, 379–384. doi: 10.3758/BF03192707

Bergeson, T. R., and Pisoni, D. B. (2004). "Audiovisual speech perception in deaf adults and children following cochlear implantation," in *The Handbook of Multisensory Processes*, eds B. E. Stein, C. Spence, and G. Calvert (Cambridge, MA: MIT Press), 749–771.

Boraston, Z., and Blakemore, S. J. (2007). The application of eye-tracking technology in the study of autism. *J. Physiol.* 581, 893–898. doi: 10.1113/jphysiol.2007.133587

Buchan, J., Parè, M., and Munhall, K. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Soc. Neurosci.* 2, 1–13. doi: 10.1080/17470910601043644

Chawarska, K., and Shic, F. (2009). Looking but not seeing: atypical visual scanning and recognition of faces in 2 and 4 year old children with autism spectrum disorder. *J. Autism Dev. Disord.* 39, 1663–1672. doi: 10.1007/s10803-009-0803-7

Cohen, J. (1973). Eta-squared and partial eta squared in fixed factor ANOVA designs. *Educ. Psychol. Meas.* 33, 107–113. doi: 10.1177/001316447303300111

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Shaeferm, H. S., Gernsbacher, M. A., Goldsmith, H. H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nat. Neurosci.* 8, 519–526. doi: 10.1038/nn1421

Dawson, G., Webb, S. J., and McPartland, J. (2005). Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. *Dev. Neuropsychol.* 27, 403–424. doi: 10.1207/s15326942dn2703_6

De Gelder, B., Vroomen, J., and Van Der Heide, L. (1991). Face recognition and lip reading in autism. *Eur. J. Cogn. Psychol.* 3, 69–86. doi: 10.1080/09541449108406220

Desjardins, R. N., Rogers, J., and Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *J. Exp. Child Psychol.* 66, 85–110. doi: 10.1006/jecp.1997.2379

Elliott, C. D. (1991). *Differential Ability Scales: Introductory and Technical Handbook*. New York: The Psychological Corporation.

Falck-Ytter, T., Fernell, E., Hedvall, A. L., Von Hofsten, C., and Gillberg, C. (2012). Gaze performance in children with autism spectrum disorder when observing communicative actions. 42, 2236–2245. doi: 10.1007/s10803-012-1471-6

Falck-Ytter, T., and von Hofsten, C. (2011). How special is social looking in ASD: a review. *Prog. Brain Res.* 189, 209–222. doi: 10.1016/B978-0-444-53884-0.00026-9

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H. P., Russo, N. N., Blanco, D., et al. (2013). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cereb. Cortex* doi: 10.1093/cercor/bht213 [Epub ahead of print].

Hobson, R. P., Ouston, J., and Lee, A. (1988). What's in a face? The case of autism. *Br. J. Psychol.* 79, 441–453. doi: 10.1111/j.2044-8295.1988.tb02745.x

Hutt, C., and Ounstead, C. (1966). The biological significance of gaze aversion with particular reference to the syndrome of infantile autism. *Behav. Sci.* 11, 346–356. doi: 10.1002/bs.3830110504

Iarocci, G., and McDonald, J. (2006). Sensory integration and the perceptual experience of persons with autism. *J. Autism Dev. Disord.* 36, 77–90. doi: 10.1007/s10803-005-0044-3

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., and Chua, R. (2010). Visual influences on speech perception in children with autism. *Autism* 14, 305–320. doi: 10.1177/1362361309353615

Irwin, J. R., Tornatore, L., Brancazio, L., and Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Dev.* 82, 1397–1403. doi: 10.1111/j.1467-8624.2011.01619.x

Jemel, B., Mottron, L., and Dawson, M. (2006). Impaired face processing in autism: fact or artifact? *J. Autism Dev. Disord.* 36, 91–106. doi: 10.1007/s10803-005-0050-5

Kleinhans, N. M., Richard, T., Sterling, L., Stegbauer, K. C., Mahurin, R., Johnson, L. C., et al. (2008). Abnormal functional connectivity in autism spectrum disorders during face processing. *Brain* 131, 1000–1012. doi: 10.1093/brain/awm334

Klin, A., Jones, W., Schultz, R., Volkmar, F. R., and Cohen, D. J. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiatry* 59, 809–816. doi: 10.1001/archpsyc.59.9.809

Lachs, L., and Pisoni, D. B. (2004). Specification of cross-modal source information in isolated kinematic displays of speech. *J. Acoust. Soc. Am.* 116, 507–518. doi: 10.1121/1.1757454

Lachs, L., Pisoni, D. B., and Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: a first report. *Ear Hear.* 22, 236–251. doi: 10.1097/00003446-200106000-00007

Lord, C., Risi, S., Lambrecht, L., Cook, E., Leventhal, B., DiLavore, P., et al. (2000). The autism diagnostic observation schedule generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* 30, 205–223. doi: 10.1023/A:1005592401947

Lord, C., Rutter, M., DiLavore, P. C., and Risi, S. (2002). *Autism Diagnostic Observation Schedule: Manual.* Los Angeles, CA: Western Psychological Services.

Lord, C., Rutter, M., and LeCouteur, A. (1994). Autism Diagnostic Interview Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* 24, 659–685. doi: 10.1007/BF02172145

MacDonald, J., Andersen, S., and Bachmann, T. (2000). Hearing by eye: how much spatial degradation can be tolerated? *Perception* 29, 1155–1168. doi: 10.1068/p3020

MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. *Percept. Psychophys.* 24, 253–257. doi: 10.3758/BF03206096

Massaro, D. W., and Bosseler, A. (2003). Perceiving speech by ear and eye: multimodal integration by children with autism. *J. Dev. Learn. Disord.* 7, 111–146.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Meltzoff, A. N., and Kuhl, P. K. (1994). "Faces and speech: intermodal processing of biologically relevant signals in infants and adults," in *The Development of Intersensory Perception: Comparative Perspectives*, eds D. J. Lewkowitz and R. Lickliter (Hillsdale, NJ: Erlbaum), 335–369.

Menard, L., Dupont, S., Baum, S. R., and Aubin, J. (2009). Perception and perception of French vowels by congenitally blind adults and sighted adults. *J. Acoust. Soc. Am.* 126, 1406–1414. doi: 10.1121/1.3158930

Mongillo, E., Irwin, J. R., Whalen, D. H., Klaiman, C., Carter, A. S., and Schultz, R. T. (2008). Audiovisual processing in children with and without autism spectrum disorders. *J. Autism Dev. Disabil.* 38, 1439–1458. doi: 10.1007/s10803-007-0521-y

Neumann, D., Spezio, M. L., Piven, J., and Adolphs, R. (2006). Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention. *Soc. Cogn. Affect. Neurosci.* 1, 194–202. doi: 10.1093/scan/nsl030

Norbury, C. F., Brock, J., Cragg, L., Einav, S., Griffiths, H., and Nation, K. (2009). Eye movement patterns are associated with communicative competence in autism spectrum disorders. *J. Child Psychol. Psychiatry* 50, 834–852. doi: 10.1111/j.1469-7610.2009.02073.x

Olejnik, S., and Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychol. Methods* 8, 434–447. doi: 10.1037/1082-989X.8.4.434

Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., and Piven, J. (2002). Visual scanning of faces in autism. *J. Autism Dev. Disord.* 32, 249–261. doi: 10.1023/A:1016374617369

Phillips, W., Baron-Cohen, S., and Rutter, M. (1992). The role of eye contact in goal detection: evidence from normal infants and children with autism or mental handicap. *Dev. Psychopathol.* 4, 375–383. doi: 10.1017/S0954579400000845

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (London, UK: Lawrence Erlbaum Associates, Ltd), 97–113.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–950. doi: 10.1126/science.7233191

Saalasti, S., Kätsryi, J., Tiippana, K., Laine-Hernandez, M., von Wendt, L., and Sams, M. (2012). Audiovisual speech perception and eye gaze behavior of adults with asperger syndrome. *J. Autism Dev. Disord.* 42, 1606–1615. doi: 10.1007/s10803-011-1400-0

Semel, E, Wiig E. H., and Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals 4 (CELF-4) Technical Manual.* San Antonio, TX: The Psychological Corporation.

Senju, A., and Johnson, M. H. (2009). Atypical eye contact in autism: models, mechanisms and development. *Neurosci. Biobehav. Rev.* 33, 1204–1214. doi: 10.1016/j.neubiorev.2009.06.001

Smith, E., and Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *J. Child Psychol.* 48, 813–821. doi: 10.1111/j.1469-7610.2007.01766.x

Speer, L. L., Cook, A. E., McMahon, W. M., and Clark, E. (2007). Face processing in children with autism. *Autism* 11, 265–277. doi: 10.1177/1362361307076925

Spezio, M. L., Adolphs, R., Hurley, R. S. E., and Piven, J. (2007). Abnormal use of facial information in high-functioning autism. *J. Autism Dev. Disord.* 37, 929–939. doi: 10.1007/s10803-006-0232-9

Sterling, L., Dawson, G., Webb, S., Murias, M., Munson, J., Panagiotidies, H., et al. (2008). The role of face familiarity in eye tracking of faces by individuals with autism spectrum disorders. *J. Autism Dev. Disord.* 38, 1666–1675. doi: 10.1007/s10803-008-0550-1

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 21. doi: 10.1121/1.1907384

Thomas, S. M., and Jordan, T. R. (2004). Contributions to oral and extraoral facial movement to visual and audiovisual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 873–888. doi: 10.1037/0096-1523.30.5.873

Volkmar, F. R., and Mayes, L. C. (1990). Gaze behavior in autism. *Dev. Psychopathol.* 2, 61–69. doi: 10.1017/S0954579400000596

Volkmar, F. R., Sparrow, S. S., Rende, R. D., and Cohen, D. J. (1989). Facial perception in autism. *J. Child Psychol. Psychiatry* 30, 591–598. doi: 10.1111/j.1469-7610.1989.tb00270.x

Williams, J. H. G., Massaro, D. W., Peel, N. J., Bosseler, A., and Suddendorf, T. (2004). Visual-auditory integration during speech imitation in autism. *Res. Dev. Disabil.* 25, 559–575. doi: 10.1016/j.ridd.2004.01.008

Woynaroski, T. G., Kwayke, L. D., Foss-Feig, J. H., Stevenson, R., Stone, W. L., and Wallace, M. (2013). Multisensory speech perception in children with autism

spectrum disorders. *J. Autism Dev. Disord.* 43, 2891–2902. doi: 10.1007/s10803-013-1836-5

Yi, A., Wong, W., and Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *J. Speech Lang. Hear. Res.* 56, 471–480. doi: 10.1044/1092-4388(2012/10-0288)

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 February 2014; accepted: 15 April 2014; published online: 08 May 2014.

*Citation: Irwin JR and Brancazio L (2014) Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. Front. Psychol. **5**:397. doi: 10.3389/fpsyg.2014.00397*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 04 November 2014
doi: 10.3389/fpsyg.2014.01198

Temporal factors affecting somatosensory–auditory interactions in speech processing

***Takayuki Ito*[1]*, *Vincent L. Gracco*[1,2] *and David J. Ostry*[1,2]**

[1] *Haskins Laboratories, New Haven, CT, USA*
[2] *McGill University, Montréal, QC, Canada*

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Martin Schürmann, University of Nottingham, UK
Donald Derrick, University of Canterbury, New Zealand

***Correspondence:**
Takayuki Ito, Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA
e-mail: taka@haskins.yale.edu

Speech perception is known to rely on both auditory and visual information. However, sound-specific somatosensory input has been shown also to influence speech perceptual processing (Ito et al., 2009). In the present study, we addressed further the relationship between somatosensory information and speech perceptual processing by addressing the hypothesis that the temporal relationship between orofacial movement and sound processing contributes to somatosensory–auditory interaction in speech perception. We examined the changes in event-related potentials (ERPs) in response to multisensory synchronous (simultaneous) and asynchronous (90 ms lag and lead) somatosensory and auditory stimulation compared to individual unisensory auditory and somatosensory stimulation alone. We used a robotic device to apply facial skin somatosensory deformations that were similar in timing and duration to those experienced in speech production. Following synchronous multisensory stimulation the amplitude of the ERP was reliably different from the two unisensory potentials. More importantly, the magnitude of the ERP difference varied as a function of the relative timing of the somatosensory–auditory stimulation. Event-related activity change due to stimulus timing was seen between 160 and 220 ms following somatosensory onset, mostly around the parietal area. The results demonstrate a dynamic modulation of somatosensory–auditory convergence and suggest the contribution of somatosensory information for speech processing process is dependent on the specific temporal order of sensory inputs in speech production.

Keywords: facial skin sensation, speech perception, speech production, electroencephalography, event-related potentials

INTRODUCTION

Multiple sensory inputs seamlessly interact in the process of speech perception. Information from a talker comes to a listener by way of the visual and auditory systems. The McGurk effect (McGurk and MacDonald, 1976) is a compelling demonstration of how the influence of audio–visual (AV) information is used in speech perceptual processing. Electrophysiological (Giard and Peronnet, 1999; Foxe et al., 2000; Molholm et al., 2002) and functional imaging studies (Macaluso et al., 2000; Calvert et al., 2001; Foxe and Simpson, 2002) have provided evidence that cortical multisensory integration can occur at early stages of cortical processing. In addition, evidence for multisensory AV processing has been identified over left central scalp which has been hypothesized to reflect sensorimotor integration (Molholm et al., 2002).

In contrast to the main focus of AV interactions, recent findings of an orofacial somatosensory influence on the perception of speech sounds suggest a potential role for the somatosensory system in speech processing. For example, air puffs to the cheek that coincide with auditory speech stimuli alter participants' perceptual judgments (Gick and Derrick, 2009). In addition, precise orofacial stretch applied to the facial skin while people listen to words, alters the sounds they hear as long as the stimulation applied to the facial skin is similar to the stimulation that normally accompanies speech production (Ito et al., 2009). Whereas these and other psychophysics experiments have examined somatosensory–auditory interactions during speech processing in behavioral terms (Fowler and Dekle, 1991), neuroimaging studies exploring the relation between multisensory inputs have been limited to AV interaction (van Atteveldt et al., 2007; Pilling, 2009; Vroomen and Stekelenburg, 2010; Liu et al., 2011).

The temporal relationship between multiple sensory inputs is one important factor for the tuning of multi-sensory interactions (Vroomen and Keetels, 2010). At a behavioral level, multiple sensory inputs are not required to arrive exactly at the same time, but some level of temporal proximity is needed to induce an interaction. In AV speech, the visual inputs influence speech perception even when the visual input leads the auditory input by as much as 200 ms (Munhall et al., 1996; van Wassenhove et al., 2007). Temporal relationships have also been examined in somatosensory–auditory interactions (see review for Occelli et al., 2011), but only in temporal perception of non-speech processing. In speech production, the temporal relationship between somatosensory inputs associated with articulatory

motion and their acoustic consequences varies. For the most part, somatosensory inputs due to articulatory motion occur in advance of acoustic output (Mooshammer et al., 2012). If the influence of somatosensation on speech perception is based on the temporal mapping between somatosensory and auditory inputs that is acquired in speech production, it is plausible then that cortical potentials may be influenced in response to the specific timing of somatosensory–auditory interactions during speech processing. Moreover, the contribution of auditory–somatosensory interactions during speech processing using speech-production-like patterns of facial skin stretch, may yield important insight into the link between speech production and perception.

In the current study, we investigate auditory and somatosensory interactions during speech processing using event-related potentials (ERPs). A robotic device was used to generate patterns of facial skin deformation similar in timing and duration to those experienced during speech production, which induces an interaction with speech sound processing (Ito et al., 2009; Ito and Ostry, 2012). We observed somatosensory–auditory interactions during speech sound processing as well as a dynamic modulation of the effects of multisensory input as a result of relative timing differences between the two sensory stimuli. ERPs using electroencephalography (EEG) have benefits for the investigation of temporal asynchronies because of their better temporal resolution in comparison with the other brain imaging techniques. The findings reveal neural correlates of multisensory temporal coding and a dynamic modulation of multisensory interaction during speech processing. The results have implications for understanding the integral link between speech production and speech perception.

MATERIALS AND METHODS

PARTICIPANTS

Eighteen native speakers of American English participated in the experiment (12 for ERP recording and 6 for the separate behavioral control test). The participants were all healthy young adults with normal hearing and all reported to be right-handed. All participants signed informed consent forms approved by the Yale University Human Investigation Committee.

EXPERIMENTAL STIMULATION AND TASK

We examined interaction effects between speech sound processing and orofacial somatosensory processing. ERPs were recorded in response to either individual or paired somatosensory and auditory stimulation. The somatosensory and auditory pairs used in the current study have been found previously to induce perceptual modulation in speech sound perception (Ito et al., 2009) and somatosensory judgment (Ito and Ostry, 2012).

A small robotic device (SenSable Technology, Phantom 1.0) applied skin stretch loads for the purpose of orofacial somatosensory stimulation. The details of the somatosensory stimulation device have been described in our previous studies (Ito et al., 2009; Ito and Ostry, 2010). Briefly, two small plastic tabs were attached bilaterally with tape to the skin at the sides of the mouth. The tabs were connected to the robotic device using monofilament and skin stretch was applied in an upward direction. Skin stretch consisted of a single cycle of a 3-Hz sinusoid with 4 N maximum force. This temporal pattern has successfully induced somatosensory ERPs in a previous study (Ito et al., 2013).

Audio stimulation was delivered binaurally through plastic tubes (24 cm) and earpieces (Etymotic Research, ER3A). We used a single synthesized speech utterance that was midway in a 10-step sound continuum between "*head*" and "*had.*" The speech continuum was created by shifting the first (F1) and the second (F2) formant frequencies in equal steps (Purcell and Munhall, 2006). The original sample sounds of "*head*" and "*had*" were produced by a male native speaker of English. These same sounds were used in a previous study demonstrating modulation of speech perception in response to facial skin stretch (Ito et al., 2009). We chose the center point of the continuum as an example of a perceptually ambiguous sound. In the current study, participants reported 68.5% of stimuli as "*head*" due to the ambiguity of the stimulus.

We used three somatosensory–auditory conditions that varied according to the time lag between the stimuli. The variations were 90 ms lead, simultaneous, and 90 ms lag of the somatosensory onset relative to the auditory onset. A 90-ms temporal asynchrony was chosen because a 90-ms somatosensory lead reliably modulated speech perception in a previous study (Ito et al., 2009). **Figure 1A** shows three temporal relationships between somatosensory and auditory stimuli (lead, lag, and simultaneous). Two unisensory conditions (somatosensory alone and

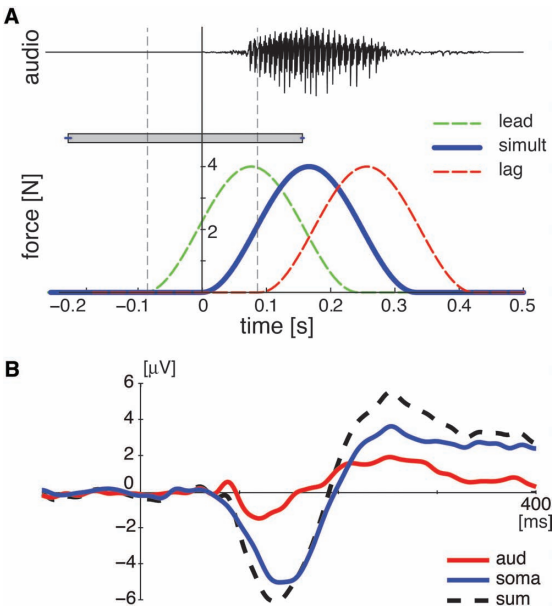


FIGURE 1 | (A) Temporal relationship in paired somatosensory–auditory stimulation. The top panel represents acoustic signal for auditory stimulation. The bottom panel represents the temporal force pattern of facial skin stretch for somatosensory stimulation. The horizontal rectangle with error bars represents the temporal range of somatosensory onset relative to auditory onset over which participants perceived somatosensory and auditory stimuli as simultaneous for the behavioral control study. **(B)** A representative example of two uni-sensory responses (aud and soma) and the sum response at Cz for the simultaneous condition. Each line represents the data averaged across all subjects.

auditory alone) were also assessed. The stimulus condition was changed every trial in random order. The inter-trial interval was varied between 1000 and 2000 ms after each response in order to avoid anticipation and habituation.

The participant's task was to indicate whether the sound they heard was "*head*" or not. The participants' response was recorded by key press. In the somatosensory alone condition, the participants were instructed to answer not "*head*" since there was no auditory stimulation. Participant judgments and the reaction time from the onset of the stimulus to the key press constituted the behavioral measures. The participants fixated their gaze on a cross without blinking in order to eliminate artifacts during ERP recording. The cross was removed every 10 trials and the participants were given a short break.

EEG ACQUISITION AND DATA PROCESSING

Recording and pre-processing

Event-related potentials were recorded from 64 electrodes (Biosemi ActiveTwo) in response to five stimulus conditions: somatosensory stimulation alone (soma), auditory stimulation alone (aud), and paired somatosensory and auditory stimulation (pair: lead, simultaneous, lag). Hundred ERPs per condition were recorded. Trials with blinks and eye movement were rejected offline on the basis of horizontal and vertical electro-oculography (over $\pm 150\ \mu V$). More than 85% of trials per condition were included in the analysis. EEG signals were filtered with a 0.5–50 Hz band-pass filter and re-referenced to the average across all electrodes. The effect of temporal manipulation was analyzed in two ways: somatosensory and auditory viewpoint depending on the alignment of the data at either the somatosensory or auditory onset. In both analyses, a single epoch was extracted in the range between −500 and 1000 ms relative to either somatosensory or auditory stimulus onset. Bias levels were adjusted using the average amplitude in the pre-stimulus interval (−200 to −100 ms).

Somatosensory analysis

The ERPs in the "pair" condition were aligned at the somatosensory onset. Neural response interactions were identified by comparing ERPs obtained using somatosensory–auditory stimulus pairs with the algebraic sum of ERPs to the unisensory stimuli presented separately by following the method applied in previous studies of somatosensory–auditory interactions (Foxe et al., 2000; Murray et al., 2005). The "sum" ERPs were calculated by summing auditory-alone potentials (aud) with an appropriate temporal shift (either 90 ms lead, 0 or 90 ms lag) with somatosensory-alone potentials (soma; see **Figure 1B** as representative example of the aud, soma, and sum ERP in the simultaneous condition). This analysis is based on the assumption that the summed ERP responses from the unisensory conditions should be equivalent to the ERP from the same stimuli presented simultaneously, if neural responses to each of the unisensory stimuli are independent (Calvert et al., 2001). Accordingly, divergence between the "sum" ERPs from the two unisensory conditions and "pair" ERPs from paired somatosensory–auditory conditions indicates non-linear interaction between the neural responses to the multisensory stimuli. Note that this approach is limited in non-linear multisensory interaction and is not sensitive to linear multisensory convergence wherein processes to two sensory modalities might interact, but not show any additional activation in electro cortical potentials.

We used the global field power (GFP) to compare the "pair" and "sum" ERPs to identify general changes in electric field strength over the entire head. GFP is the root mean square computed over the average-referenced electrode values at a given instant in time (Lehmann and Skrandies, 1980; Murray et al., 2008). GFP is equivalent to the spatial standard deviation of the scalp electric field, and yields larger values for stronger fields. The use of a global measure was in part motivated by the desire to minimize observer bias that can follow from analyses restricted to specific selected electrodes. We determined GFP amplitude using a temporal window that shows the changes in this measure over the course of the response. The corresponding temporal intervals were determined based on our observation across the three "pair" conditions described in Section "Results." ERP comparisons at the representative electrodes (Fz, Cz, and Pz) follow. These electrodes were chosen in order to sample the whole map, irrespective of asymmetry and minimizing the number of comparisons.

A 60-ms time window was chosen for the statistical analysis of the GFP amplitude and of the ERP amplitude at the representative electrodes (Fz, Cz, and Pz). In the GFP analysis, we used repeated measures two-way ANOVA to test for differences related to the relative timing of the responses for the somatosensory and auditory stimulation (90 ms lead and lag, and simultaneous) and for the difference between the "sum" of the two unisensory ERPs and "pair" somatosensory–auditory ERP. We also applied repeated measures three-way ANOVA to three electrodes (Fz, Cz, and Pz).

We also compared the topographic map differences between "sum" and "pair" ERPs across the three stimulus timing conditions (lead, lag, and simultaneous). A difference in amplitude was obtained by subtracting "sum" ERPs from "pair" ERPs at each electrode. As an index of topographic difference between the two electric fields, a global dissimilarity measures (DISS) was used (Lehmann and Skrandies, 1980). This parameter is computed as the square root of the mean of the squared difference between the potentials measured at each electrode (versus the average reference), each of which is first scaled to unitary strength by dividing by the instantaneous GFP. This value can range from 0 to 2, where 0 indicates topographic homogeneity and 2 indicates topographic inversion (Murray et al., 2008).

Auditory analysis

The ERPs in the "pair" condition were aligned at auditory onset. We reconstructed auditory-like potentials by subtracting somatosensory potentials (soma) from the "pair" potentials at the corresponding temporal shift in each condition, instead of applying the sum of two uni-sensory conditions as done in the somatosensory analysis. Our rationale is that since the mechanism of auditory ERPs in speech processing is well established (e.g., Näätänen and Picton, 1987; Martin et al., 2008 for review), comparing auditory-like potentials in the "pair" condition with the typical auditory ERP (aud) is a way to evaluate the potential multisensory interaction effect. As in the analyses using the algebraic sum described above, we expected that "pair" ERPs with the removal of the somatosensory potentials would be equivalent

to the auditory-alone ERP, if neural responses to each of the unisensory stimuli are independent. The subtracted potential should be different from the auditory responses if there is a non-linear interaction. The ERP results were also compared with participants' behavioral performance, that is, the probability that the stimulus was identified as "head" during the test as mentioned later.

We focused on the first negative peak (N1) and the following positive peak (P2) at Fz and Cz because as a general tendency the maximum amplitude of the auditory ERP is observed at these electrodes and this was true of the current responses. Note that the negative peak and positive peak do not mean negative or positive value but the direction of electrical deviation, and hence N1 can be a positive value as long as it is going in a negative direction (i.e., Ostroff et al., 1998). A 60-ms time window was used to calculate the response amplitude. The analysis window was centered at the ERP peak location for each participant and each condition. The peaks associated with N1 and P2 were identified in the time periods (100–200 ms for N1 and 200–300 ms for P2) following stimulus onset. Repeated measures ANOVA was applied to assess differences in the four conditions (three "pair" potentials and one auditory potential). Pairwise comparisons with Bonferroni correction followed.

BEHAVIORAL PERFORMANCE

Behavioral performance was evaluated using reaction time and judgment probability separately. Reaction time was calculated as the period between auditory onset and the behavioral response (key press for the speech sound identification). Repeated measures ANOVA was used to assess differences in reaction time across five conditions: three "pair" and two unisensory conditions. We also calculated the probability that the participant classified the sound as "head." The somatosensory alone condition was not included in this analysis. Note that in more than 95% of somatosensory trials participants responded not "head" as instructed. Repeated measures ANOVA was used to compare judgment measures across conditions.

We also examined the extent to which the perceptual judgments were correlated with ERP amplitude change that were observed in response to changes in the relative timing of somatosensory–auditory stimulation. The correlation analysis was carried out between the participants' judgment probability and the auditory ERP amplitude obtained when the somatosensory response was subtracted from "pair" responses and auditory-alone response. For the purpose of this analysis, both variables were transformed into z-scores in order to remove differences in amplitude variability between individuals. The analysis was applied independently at each electrode and for each response peak. Note that we did not apply correlation analysis to the data aligned at somatosensory onset because ERPs in each "pair" condition during a specific period relative to somatosensory onset represent different time periods in terms of auditory processing.

BEHAVIORAL CONTROL

As a separate control, simultaneity judgments were obtained in order to determine whether participants perceived the temporal difference (simultaneous, 90 ms lead and lag) between somatosensory and auditory stimuli as simultaneous. We assessed the perceived temporal range of somatosensory onset relative to auditory onset when both stimuli were presented simultaneously. Six individuals participated in the test. The participants were presented with auditory and somatosensory stimulation and asked to report if stimuli were simultaneous or not. The test started with two initial values of somatosensory stimulation relative to auditory onset: (1) 300 ms lead and (2) 300 ms lag. The lead and lag conditions were alternated in random order. The temporal difference between the somatosensory and auditory stimulations was reduced based on participants' response according to the Parameter Estimation by Sequential Testing (PEST) procedure (Macmillan and Creelman, 2004) until they reached a threshold level to detect the somatosensory–auditory stimulations as simultaneous or not.

RESULTS

SOMATOSENSORY-ALIGNED POTENTIALS

We first examined whether the timing difference between the sensory stimulation conditions induced changes in GFP. **Figure 2A** shows the GFP as the timing of stimulation varied (lead, lag, and simultaneous). The red thick line shows the GFP for the paired sensory condition and the black dashed line shows the sum of the two unisensory conditions (soma + aud). The data are aligned at somatosensory onset. The arrows represent the auditory onsets. The vertical dotted lines are the temporal intervals used to assess differences between conditions as a result of the timing of stimulation. Two empirically determined intervals were used to assess stimulus-timing effects after the end of the transient phase of the GFP change. The first interval is between 160 and 220ms after the somatosensory stimulus onset and the second is between 220 and 280ms. We found two pattern of differences in the target intervals respectively. For the first interval, the response amplitude difference between the "pair" and "sum" signals changed as a function of stimulus timing. In the lead condition, the "pair" GFP was marginally greater than the "sum" of the individual GFPs. The sign of the difference was reversed in the same and lag conditions. The difference in lag condition was greater than in the simultaneous condition. These amplitude differences are summarized in left panel of **Figure 2B**. Repeated measures ANOVA indicated reliable change across the three temporal conditions [$F(2,22) = 7.76$, $p < 0.01$]. For the second interval, the "pair" GFP was consistently smaller than the "sum" GFP regardless of the timing condition. These amplitude differences are summarized in the right panel of **Figure 2B**. Repeated measures ANOVA revealed that GFP in "pair" response was reliably smaller than GFP in the sum of unisensory responses [$F(1,11) = 6.81$, $p < 0.03$]. Note that there was a difference between sum and pair conditions before the transient phase (up to 80 ms after somatosensory onset). This is most likely due to an added effect of noise in the summed condition since this difference was not present for each component individually (see **Figure 2C**). Overall, these results suggest timing-related and timing-independent processing associated with separate stages of the evoked response.

We further investigated the trend observed in the GFP by examining the response patterns at individual electrodes. At the first target interval, that is, in the interval when

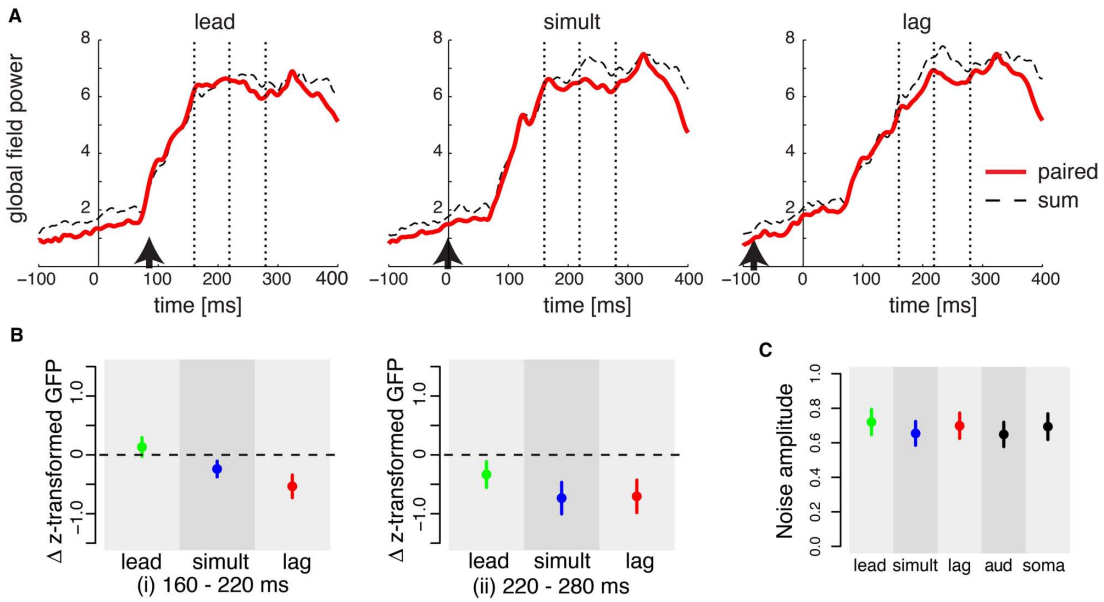


FIGURE 2 | (A) Global field power (GFP) for somatosensory and auditory stimulation in the three timing conditions (lead, simultaneous, and lag). Each line represents the data averaged across all subjects. The vertical dotted lines define two intervals (i) 160–220 ms and (ii) 220–280 ms after somatosensory onset in which differences between "pair" and "sum" responses are assessed. Arrows represents auditory onset. **(B)** Difference in z-transformed GFP amplitude in the temporal periods of (i) and (ii) in panel **(A)**. Error bars give standard errors across participants. Each color represents a different experimental condition of relative timing difference. **(C)** GFP noise level in the period before the stimulation (−200 to −100 ms) in all conditions. Error bars represent the standard error across all subjects.

somatosensory–auditory activation changed in a manner dependent on the relative timing of paired stimulation, the topographic configuration varied according to the relative timing of the stimulation (lead, simultaneous, and lag). **Figure 3** shows topographic maps of the mean differences in amplitude between "pair" and "sum" responses. In the lead case, most of the differences were in the positive direction and were observed in the parietal electrodes. Similar differences in the parietal electrodes were also seen in the simultaneous condition, although the amplitude of the difference was smaller compared to the lead condition. In the lag condition, no difference was observed in the parietal electrodes, but several frontal electrodes showed a positive difference.

The similarity of the topographic configuration was assessed using global dissimilarity as a quantitative measure (DISS, review in Murray et al., 2008). DISS values indicated that the lead and lag conditions were topographically inverted (DISS = 1.41). On the other hand, lead and simultaneous conditions were moderately homogeneous (DISS = 0.82). The similarity between the simultaneous and lag conditions was not remarkable (DISS = 1.10), suggesting that the topography for the simultaneous condition was intermediate between the lead and lag conditions. The inverted topographic configuration between the lead and lag condition suggests that the topographic configuration was altered in conjunction with the timing differences between somatosensory and auditory stimulation onsets.

Event-related potential amplitude difference in the second target interval (220–280 ms after somatosensory onset) showed consistent change across the three "pair" condition in terms of GFP (**Figure 2B**, right panel). The topographic configuration

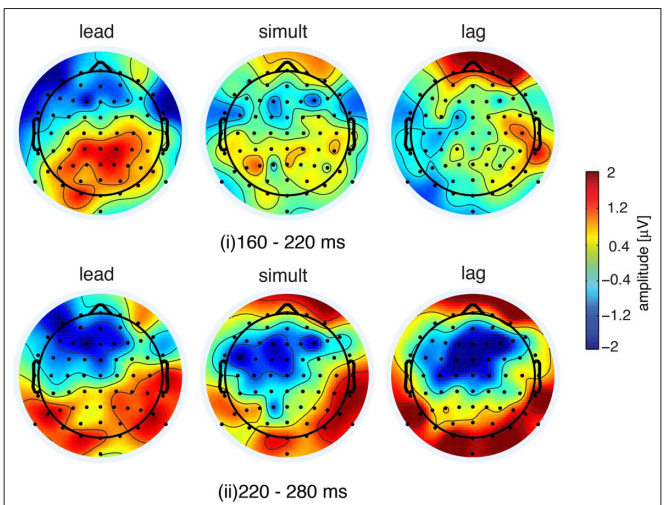


FIGURE 3 | Topographic maps of differences in event-related potentials between "pair" and "sum" conditions. The top three panels represent the difference in the interval 160–220 ms and the bottom represents the difference at 220–280 ms after somatosensory onset.

during the period 220–280 ms was quite similar across three conditions (**Figure 3**, lower panels). We found that a broad range of frontal areas showed a reduction of "pair" responses in comparison to "sum" in all three temporal conditions, as was observed in GFP (**Figure 2**). Global dissimilarity for all three conditions is comparatively homogenous [DISS = 0.64 (lead and simultaneous), 0.49 (simultaneous and lag), and 0.70 (lead

and lag)], suggesting that the amplitude reduction was present regardless of stimulus timing.

Temporal patterns of ERP in representative electrodes (Fz, Cz, and Pz) are shown in **Figure 4**. As indicated in the GFP analysis, two patterns of change across three stimulus conditions were observed in the two temporal intervals respectively. In the first interval between 160 and 220 ms, repeated measure three-way ANOVA showed a reliable interaction effect across timing (lead, simultaneous, and lag), condition (pair and sum), and electrodes (Fz, Cz, and Pz) [$F(4,44) = 3.175$, $p < 0.03$]. In a more detailed analysis with Bonferroni correction, the difference in Pz amplitude between "pair" and "sum" ERPs changed as a function of the three stimulation conditions [$F(2,22) = 5.99$, $p < 0.03$], but there was no change and no difference in the other two electrodes [Fz: $F(2,22) = 1.73$, $p > 0.6$; Cz: $F(2,22) = 1.00$, $p = 1.0$]. In contrast, in the second interval between 220 and 280 ms there was a reliable interaction effect between experimental condition (pair and sum) and electrodes (Fz, Cz, and Pz) [$F(2,22) = 9.812$, $p < 0.001$]. Following Bonferroni correction, ERP amplitude at Fz and Cz in the "pair" condition was also consistently smaller than the "sum" ERP amplitude in the three stimulation conditions [Fz: $F(1,11) = 8.32$, $p < 0.05$; Cz: $F(1,11) = 12.24$, $p < 0.02$], but there was no difference at Pz [$F(1,11) = 0.87$, $p = 1.0$].

AUDITORY-ALIGNED POTENTIALS

While we found a dynamical modulation of the ERP response in the context of somatosensory processing, it is difficult to interpret this change relative to speech processing since the observed modulation is not directly related to the timing of auditory processing. In order to compare paired auditory and somatosensory processing with that involved in speech perceptual processing, we examined these paired effects in relation to auditory-related processing on its own. We extracted auditory-related responses in the various paired conditions by subtracting the somatosensory-alone response from that obtained in the "pair" conditions. The logic is that if there is a non-linear interaction between somatosensory and auditory processing, the response after the subtraction should be different from the auditory alone response.

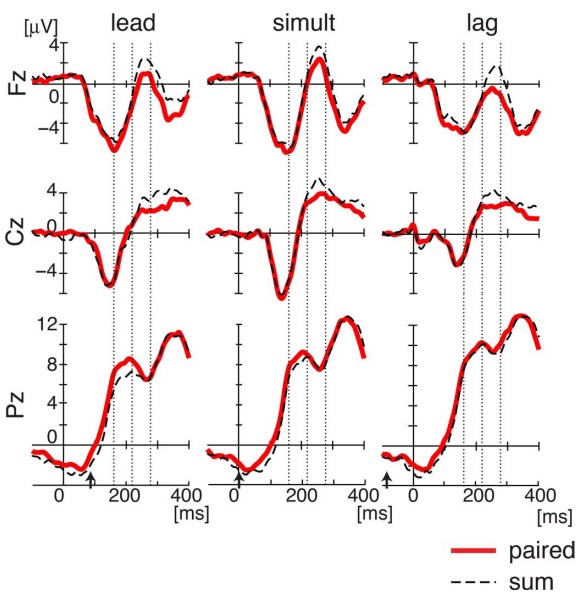


FIGURE 4 | Event-related potentials for combined somatosensory and auditory stimulation in the three timing conditions (lead, simultaneous, and lag) at Fz, Cz, and Pz. Each line represents the data averaged across all subjects. Bias levels were adjusted using the average amplitude in the pre-stimulus interval (−200 to −100 ms). As in GFP analysis, the vertical dotted lines define two intervals (i) 160–220 ms and (ii) 220–280 ms after somatosensory onset in which differences between "pair" and "sum" responses are assessed. Arrows represents auditory onset.

For this analysis, all of the data were aligned at auditory onset. The subtracted potentials and the auditory-alone potentials showed a typical N1–P2 pattern with the first negative peak (N1) between 100 and 200 ms after auditory onset followed by a second positive peak (P2) between 200 and 300 ms (see **Figure 5A**). The maximum response was observed along mid-line electrodes near Cz (vertex electrode).

The peak amplitude at the Cz and Fz electrodes was quantified using 60-ms temporal window in each of the three "pair" timing conditions (lead, simultaneous, and lag) and for the auditory response alone (**Figure 5B**). Each color represents a different condition. Error bars represent the standard error across participants. Two way-repeated measure ANOVA (timing condition × electrodes) yielded reliable differences across timing condition (lead, lag, and simultaneous) in N1 [$F(3,77) = 3.056$, $p < 0.05$] and in P2 [$F(3,77) = 6.18$, $p < 0.001$]. By looking at ERPs in each individual electrode, we found a consistent N1 response at Cz in all four conditions (lead, simultaneous, lag, and auditory). The peak amplitudes were not statistically different for the four conditions [$F(3,33) = 0.122$, $p > 0.9$]. The peak amplitude of the P2 response showed a graded change according to the stimulus timing (lead, simultaneous, and lag), although the change was statistically marginal as follows. Whereas repeated measures one-way ANOVA showed reliable difference across the four conditions [$F(3,33) = 3.82$, $p < 0.04$], *post hoc* testing did not reveal any reliable paired comparisons.

In contrast, a reliable change was observed at Fz electrode in both N1 and P2 amplitude (see right two panels in **Figure 5B**). N1 responses at Fz were reliably different across the four conditions [$F(3,33) = 5.95$, $p < 0.02$]. *Post hoc* tests with Bonferroni correction showed that the lead condition was reliably different from lag condition ($p < 0.02$) and marginally different from the auditory response ($p = 0.06$). P2 response were also reliably different across conditions [$F(3,33) = 4.80$, $p < 0.02$]. Comparing the auditory alone responses with the other "pair" condition yielded a reliable difference from the lead condition ($p < 0.05$) and a marginal difference from simultaneous condition ($p = 0.10$). The difference for the lag condition was not reliable ($p > 0.9$). Overall, the results reveal that auditory ERPs show a change when combined with temporally offset somatosensory stimulation. The largest change occurs when somatosensory stimulation lead for the speech sound. On the other hand, when somatosensory stimulation lags speech onset, the amplitude of the auditory potentials are no different from the potentials for auditory stimulation alone.

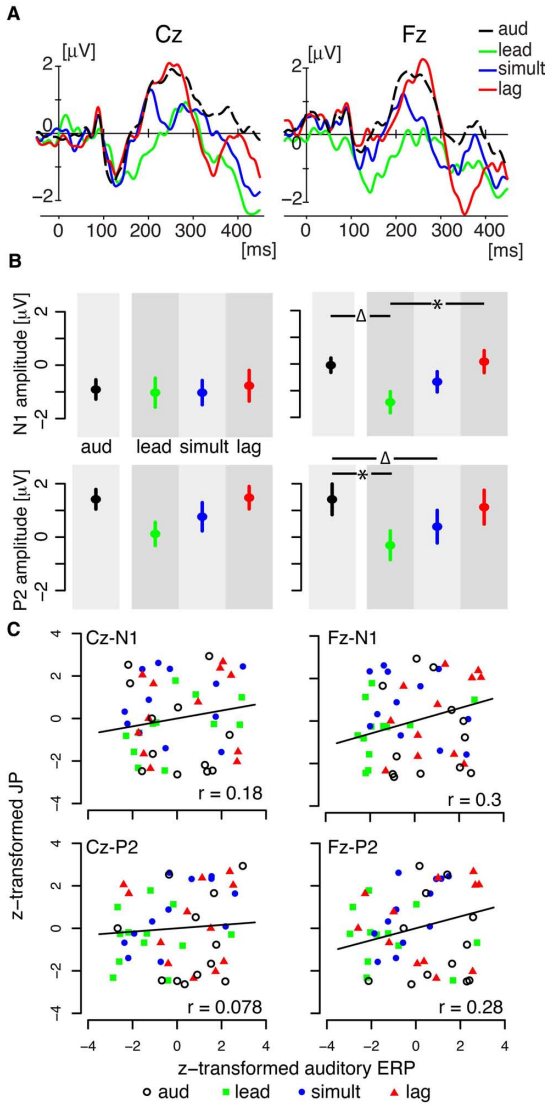


FIGURE 5 | (A) Temporal responses of auditory event-related potentials at two mid-frontal electrodes (Fz and Cz). Each line represents the data averaged across all subjects. Each color corresponds to the following four conditions. "aud" represents the amplitude of auditory alone condition. "lead," "simultaneous," and "lag" represent the amplitudes in the corresponding three "pair" conditions after the removal of somatosensory potentials. Bias levels were adjusted using the average amplitude in the pre-stimulus interval (−200 to −100 ms). **(B)** Peak amplitude of N1 and P2 auditory event-related potentials at two mid-frontal electrodes. The left two panels are for Cz, and the right two panels for Fz. Error bars give standard errors across participants. The symbol "*" represents reliable difference ($p < 0.05$) and "Δ" represents marginal difference ($p \leq 0.1$). **(C)** Correlation plots between peak amplitude of ERP and judgment probability for the behavioral task (48 data points from 12 subjects × 4 conditions).

BEHAVIORAL PERFORMANCE

We also examined the behavioral results and their relationship with EEG activity. There was no reliable change of judgment probability in the three paired conditions in comparison to the auditory alone condition [$F(3,33) = 1.128$, $p > 0.3$]. Correlation analysis showed that the judgment probabilities in the four conditions were reliably correlated with N1 amplitude at Fz ($r = 0.3$, $p < 0.05$) and marginally correlated with P2 amplitude at Fz ($r = 0.28$, $p = 0.05$). The peak amplitude of N1 and P2 at Cz were not reliably correlated with the judgment probabilities (N1: $r = 0.18$, $p > 0.2$; P2: $r = 0.078$, $p > 0.6$). **Figure 5C** shows the correlation plots in each combination of N1 and P2 responses at Cz and Pz. Thus, overall, although the magnitude of the correlation was relatively low, the results suggest that perceptual modulation as measured behaviorally may be represented to some degree in the cortical response at Fz.

Reaction times across the five conditions: three "pair" conditions and two uni-sensory conditions "soma" and "aud" were evaluated. We did not find any reliable differences across all five conditions [$F(4,44) = 0.532$, $p > 0.70$]. This is inconsistent with typical responses due to multisensory stimulation conditions. Reaction time to respond to stimuli typically becomes shorter when two sensory modalities were stimulated simultaneously than when single sensory modalities are stimulated. The difference from the typical multisensory reaction may presumably be because the current task involved identification only.

BEHAVIORAL CONTROL

In order to examine if the time differences between stimuli in the "pair" condition are perceived as simultaneous, we obtained threshold values at which the participants determined whether or not the somatosensory and auditory stimulations were simultaneous. The average threshold times for perception of simultaneity were 210.6 ± 3.1 ms lead and 148.0 ± 3.9 lag of the somatosensory onset relative to the auditory onset (see rectangle and error bar in **Figure 1A**). The results indicates that the 90 ms difference used in the current ERP recording is in a range where stimuli are perceived to be simultaneous and suggests the participants did not perceive a difference in stimulus timing in any of the three "pair" conditions.

DISCUSSION

This study assessed the neural correlate of the temporal interaction between orofacial somatosensory and speech sound processing. The cortical activity associated with orofacial somatosensory–auditory interaction was quantified using ERPs. We found two types of non-linear interactions between somatosensory and auditory processing. One form of sensory interaction was dependent on the relative timing of the two sensory stimuli. The other was constant regardless of stimulus timing. The two interactions were observed at different electrodes sites: the stimulus timing interaction was recorded over parietal electrodes and the non-stimulus timing interaction was observed over the frontal electrodes. From the viewpoint of auditory processing, we also found a graded change in ERP amplitudes that was dependent on the relative timing of stimuli for auditory processing. The results demonstrate clear multisensory convergence and suggest a dynamic modulation of these particular (somatosensory–auditory) interactions during speech processing.

It is important to note that in the previous psychophysical study demonstrating an interaction between speech sound processing and orofacial stimulation, perceptual judgments were

influenced by speech-production-like patterns of facial skin stretch (Ito et al., 2009). The current finding showing the largest amplitude change in the multisensory evoked response occurring with a somatosensory lead is also consistent with speech production-like patterning affecting cortical evoked potentials. Auditory input from self-generated speech is always preceded by articulatory motion that generates somatosensory input in advance of the acoustic signal. Interestingly, Möttönen et al. (2005) showed that simple lip tapping during speech perceptual processing did not change magnetoenchalographic evoked potentials. It appears, consistent with the previous psychophysical study (Ito et al., 2009), that the influence of somatosensory stimulation on speech perceptual processing may be dependent on a functional relationship between the somatosensory characteristics of the stimulation and those accompanying speech production. Hence, somatosensory inputs that are similar to those experienced in speech production can interact effectively with speech sound processing and the interaction is reflected in cortical potential changes.

The timing of sensory stimulation is a key factor in multisensory interaction. The effective time-window for multisensory integration is known to be as long as 200 ms (Meredith et al., 1987; van Wassenhove et al., 2007). At a behavioral level, this is consistent with the results of our control test in which the participants perceived the skin stretch perturbation and the speech sound "*head*" as simultaneous in a comparable temporal range. Although the neural correlates of AV interaction including that involving speech stimuli has been previously investigated (Pilling, 2009; Vroomen and Stekelenburg, 2010; Liu et al., 2011), the temporal range was larger than 200 ms, and hence it is not known the extent to which multisensory interactions occur at shorter temporal asynchronies. In the present study, dynamical modulation at an electrocortical level was found at a range of 100 ms. The current finding suggests cortical processing is sensitive to temporal factors even within the time range at which events are behaviorally judged simultaneous.

In AV speech, the effective temporal range between auditory and visual stimulus onsets for effective multisensory interaction is asymmetric in terms of onset timing. While AV speech phenomena, such as the McGurk effect is induced with up to a 240-ms of visual lead, while for visual lag the time window is much shorter (up to 40 ms; Munhall et al., 1996; van Wassenhove et al., 2007). Our ERP findings may be comparable. N1 and P2 potential amplitudes in the 90 ms somatosensory lag relative to auditory onset were not different from those in the auditory alone response, whereas the lead and simultaneous condition showed a difference between the "pair" and "sum" responses, indicating that the somatosensory lead condition has affected audio processing, but not in the lag condition. This can probably be attributed to the temporal relationship between orofacial somatosensory inputs and acoustic output in speech production, since articulatory motion mostly precedes acoustic output in speech production (e.g., Mooshammer et al., 2012).

In addition to the differential cortical response dependent on the asynchrony of the somatosensory–auditory stimulation, we also found a consistent reduction in the cortical response independent of the asynchrony of the stimulation in the later period (220–280 ms after somatosensory onset). Interestingly this reduction was seen only in the somatosensory analysis suggesting that this later period of somatosensory processing consistently interacts with the auditory input regardless of the timing of auditory processing. While such an obligatory multisensory interaction is a plausible interpretation, the reduction could be influenced by non-stimulus-specific factors such as changing attentional demands. However, the use of stimulus averaging over a number of the responses time-locked to the onset of specific stimulus would most likely eliminate or minimize any non-stimulus-specific effects. Consequently, the possibility of a non-specific effect to explain the consistent reduction in the cortical response is unlikely.

Two different patterns of activation were observed depending on the asynchrony of the stimulation and the specific time post-stimulation onset. The asynchrony dependent modulation was observed in a period between 160–220 ms after somatosensory onset mostly at parietal electrodes. In contrast, frontal electrodes showed a consistent multisensory change in activation in all three temporal conditions during the 220–280 ms period after somatosensory onset. Since multiple subcortical and cortical locations are involved in auditory–somatosensory interactions in non-speech processing (Stein and Meredith, 1993; Foxe et al., 2002; Lütkenhöner et al., 2002; Fu et al., 2003; Kayser et al., 2005; Murray et al., 2005; Schürmann et al., 2006; Shore and Zhou, 2006; Lakatos et al., 2007; Beauchamp et al., 2008), the present results reflect the contribution of different and distributed cortical sites in the somatosensory–auditory interaction during speech processing. Given that the parietal area and planum temporale is considered as a center of auditory–motor integration (Hickok et al., 2011; Tremblay et al., 2013), the parietal site may also be important for the temporal processing between somatosensory and auditory inputs.

While we found a reliable difference for the timing manipulation in the ERP changes, there was no such reliable result in the behavioral measure. It appears that the nervous system is sensitive to timing differences in the relatively early phase of speech processing (N1 and P2), but this difference may be independent of the ability to identify such differences. This is not surprising given the additional cognitive process involved in perceptual judgments. It appears that a more sensitive task is required to detect these subtle timing differences behaviorally. The current experimental design was optimized for identifying the influence of sensory input on cortical potentials rather than on cognitive decisions.

A possible neural pathway for somatosensory influence on speech perception is currently unknown. For language processing, the posterior inferior frontal gyrus (Broca's area) is known to contribute to speech perception, and hence the neural connections between the prefrontal and the temporal areas associated with auditory processing have been well documented (Geschwind, 1970; Margulies and Petrides, 2013). Given a connection between the premotor and somatosensory cortex, and the premotor cortex and Broca's region (Margulies and Petrides, 2013), somatosensory inputs may influence auditory processing via the prefrontal gyrus and the premotor cortex. On the other hand, studies of non-speech processing have shown that somatosensory inputs directly affect lower levels of auditory

processing in the auditory cortex and/or the surrounding areas (Foxe et al., 2002; Lütkenhöner et al., 2002; Fu et al., 2003; Kayser et al., 2005; Murray et al., 2005; Schürmann et al., 2006; Lakatos et al., 2007; Beauchamp et al., 2008). These non-speech studies have shown a reciprocal somatosensory–auditory interaction that is independent of motor involvement, in particular, change in second somatosensory cortex in response to sound and changes in auditory cortex in response to somatosensory stimulation. In addition, there are somatosensory–auditory interactions in subcortical areas (Stein and Meredith, 1993; Shore and Zhou, 2006). These results suggest a tight linkage and direct neural connection between somatosensory and auditory system that is separate from a motor–auditory connection mentioned above. The current findings show a dynamic modulation of the effects of somatosensory and auditory stimuli at the electrodes over the parietal region. This is consistent with the idea of direct access of somatosensory inputs to the auditory system. Since speech and non-speech sounds are processed differently in the brain (Kozou et al., 2005; Möttönen et al., 2006), it is unclear whether pathways associated with somatosensory–auditory interactions in non-speech processes are also involved in speech processing. Further investigation is required.

The linkage between speech production and perception processing has been a topic of interest for over five decades (Liberman et al., 1967; Diehl et al., 2004; Schwartz et al., 2012). Whereas the idea has been previously tested from the viewpoint of speech production and motor function (Fadiga et al., 2002; Watkins et al., 2003; Wilson et al., 2004; Meister et al., 2007; D'Ausilio et al., 2009; Möttönen and Watkins, 2009), the role of somatosensory function in speech perception has been overlooked. Previous psychophysical findings have showed that orofacial somatosensory inputs can influence speech processing (Ito et al., 2009). The current findings further suggest that somatosensory stimulation has access to cortical areas associated with speech processing. One intriguing possibility is that somatosensory information may be an important component in establishing the neural representations for both speech production and speech perception. Further investigation of the manner in which orofacial somatosensation modulates speech perceptual processing may provide some important clues to understanding the development of the linkage between speech perception and production.

ACKNOWLEDGMENTS

We thank Alexis R. Johns and Joshua H. Coppola for data recording and processing. This work was supported by the National Institute on Deafness and Other Communication Disorders Grants (R03DC009064 and R01DC012502), and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Beauchamp, M. S., Yasar, N. E., Frye, R. E., and Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *Neuroimage* 41, 1011–1020. doi: 10.1016/j.neuroimage.2008.03.015

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

Calvert, G. A., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001). Detection of audio–visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage* 14, 427–438. doi: 10.1006/nimg.2001.0812

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.* 55, 149–179. doi: 10.1146/annurev.psych.55.090902.142028

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402. doi: 10.1046/j.0953-816x.2001.01874.x

Fowler, C. A., and Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform*. 17, 816–828. doi: 10.1037/0096-1523.17.3.816

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., and Schroeder, C. E. (2000). Multisensory auditory–somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Brain Res. Cogn. Brain Res.* 10, 77–83. doi: 10.1016/S0926-6410(00)00024-0

Foxe, J. J., and Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. *Exp. Brain Res.* 142, 139–150. doi: 10.1007/s00221-001-0906-7

Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., et al. (2002). Auditory–somatosensory multisensory processing in auditory association cortex: an fMRI study. *J. Neurophysiol.* 88, 540–543. doi: 10.1152/jn.00694.2001

Fu, K. M. G., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., et al. (2003). Auditory cortical neurons respond to somatosensory stimulation. *J. Neurosci.* 23, 7510–7515.

Geschwind, N. (1970). The organization of language and the brain. *Science* 170, 940–944. doi: 10.1126/science.170.3961.940

Giard, M. H., and Peronnet, F. (1999). Auditory–visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 11, 473–490. doi: 10.1162/089892999563544

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462, 502–504. doi: 10.1038/nature08572

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Ito, T., Johns, A. R., and Ostry, D. J. (2013). Left lateralized enhancement of orofacial somatosensory processing due to speech sounds. *J. Speech Lang. Hear. Res.* 56, 1875–1881. doi: 10.1044/1092-4388(2013/12-0226)

Ito, T., and Ostry, D. J. (2010). Somatosensory contribution to motor learning due to facial skin deformation. *J. Neurophysiol.* 104, 1230–1238. doi: 10.1152/jn.00199.2010

Ito, T., and Ostry, D. J. (2012). Speech sounds alter facial skin sensation. *J. Neurophysiol.* 107, 442–447. doi: 10.1152/jn.00029.2011

Ito, T., Tiede, M., and Ostry, D. J. (2009). Somatosensory function in speech perception. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1245–1248. doi: 10.1073/pnas.0810063106

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron* 48, 373–384. doi: 10.1016/j.neuron.2005.09.018

Kozou, H., Kujala, T., Shtyrov, Y., Toppila, E., Starck, J., Alku, P., et al. (2005). The effect of different noise types on the speech and non-speech elicited mismatch negativity. *Hear. Res.* 199, 31–39. doi: 10.1016/j.heares.2004.07.010

Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., and Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53, 279–292. doi: 10.1016/j.neuron.2006.12.011

Lehmann, D., and Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalogr. Clin. Neurophysiol.* 48, 609–621. doi: 10.1016/0013-4694(80)90419-8

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Liu, B., Jin, Z., Wang, Z., and Gong, C. (2011). The influence of temporal asynchrony on multisensory integration in the processing of asynchronous audio–visual stimuli of real-world events: an event-related potential study. *Neuroscience* 176, 254–264. doi: 10.1016/j.neuroscience.2010.12.028

Lütkenhöner, B., Lammertmann, C., Simões, C., and Hari, R. (2002). Magnetoencephalographic correlates of audiotactile interaction. *Neuroimage* 15, 509–522. doi: 10.1006/nimg.2001.0991

Macaluso, E., Frith, C. D., and Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science* 289, 1206–1208. doi: 10.1126/science.289.5482.1206

Macmillan, N. A., and Creelman, C. D. (2004). *Detection Theory: A User's Guide*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Margulies, D. S., and Petrides, M. (2013). Distinct parietal and temporal connectivity profiles of ventrolateral frontal areas involved in language production. *J. Neurosci.* 33, 16846–16852. doi: 10.1523/JNEUROSCI.2259-13.2013

Martin, B. A., Tremblay, K. L., and Korczak, P. (2008). Speech evoked potentials: from the laboratory to the clinic. *Ear Hear.* 29, 285–313. doi: 10.1097/AUD.0b013e3181662c0e

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J. Neurosci.* 7, 3215–3229.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Res. Cogn. Brain Res.* 14, 115–128. doi: 10.1016/S0926-6410(02)00066-6

Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E., and Tiede, M. (2012). Bridging planning and execution: temporal planning of syllables. *J. Phon.* 40, 374–389. doi: 10.1016/j.wocn.2012.02.002

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage* 30, 563–569. doi: 10.1016/j.neuroimage.2005.10.002

Möttönen, R., Järveläinen, J., Sams, M., and Hari, R. (2005). Viewing speech modulates activity in the left SI mouth cortex. *Neuroimage* 24, 731–737. doi: 10.1016/j.neuroimage.2004.10.011

Möttönen, R., and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825. doi: 10.1523/JNEUROSCI.6018-08.2009

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. *Percept. Psychophys.* 58, 351–362. doi: 10.3758/BF03206811

Murray, M. M., Brunet, D., and Michel, C. M. (2008). Topographic ERP analyses: a step-by-step tutorial review. *Brain Topogr.* 20, 249–264. doi: 10.1007/s10548-008-0054-5

Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., et al. (2005). Grabbing your ear: rapid auditory–somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cereb. Cortex* 15, 963–974. doi: 10.1093/cercor/bhh197

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Occelli, V., Spence, C., and Zampini, M. (2011). Audiotactile interactions in temporal perception. *Psychon. Bull. Rev.* 18, 429–454. doi: 10.3758/s13423-011-0070-4

Ostroff, J. M., Martin, B. A., and Boothroyd, A. (1998). Cortical evoked response to acoustic change within a syllable. *Ear Hear.* 19, 290–297. doi: 10.1097/00003446-199808000-00004

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J. Speech Lang. Hear. Res.* 52, 1073–1081. doi: 10.1044/1092-4388(2009/07-0276)

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

Schürmann, M., Caetano, G., Hlushchuk, Y., Jousmäki, V., and Hari, R. (2006). Touch activates human auditory cortex. *Neuroimage* 30, 1325–1331. doi: 10.1016/j.neuroimage.2005.11.020

Schwartz, J. L., Basirat, A., Menard, L., and Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguist.* 25, 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Shore, S. E., and Zhou, J. (2006). Somatosensory influence on the cochlear nucleus and beyond. *Hear. Res.* 216–217, 90–99. doi: 10.1016/j.heares.2006.01.006

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, MA: MIT Press.

Tremblay, P., Deschamps, I., and Gracco, V. L. (2013). Regional heterogeneity in the processing and the production of speech in the human planum temporale. *Cortex* 49, 143–157. doi: 10.1016/j.cortex.2011.09.004

van Atteveldt, N. M., Formisano, E., Blomert, L., and Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cereb. Cortex* 17, 962–974. doi: 10.1093/cercor/bhl007

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884. doi: 10.3758/APP.72.4.871

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596. doi: 10.1162/jocn.2009.21308

Watkins, K. E., Strafella, A. P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41, 989–994. doi: 10.1016/S0028-3932(02)00316-0

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 24 January 2014; accepted: 04 October 2014; published online: 04 November 2014.

*Citation: Ito T, Gracco VL and Ostry DJ (2014) Temporal factors affecting somatosensory–auditory interactions in speech processing. Front. Psychol. **5**:1198. doi: 10.3389/fpsyg.2014.01198*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 10 July 2014
doi: 10.3389/fpsyg.2014.00656

Temporal dynamics of sensorimotor integration in speech perception and production: independent component analysis of EEG data

David Jenson[1], Andrew L. Bowers[2], Ashley W. Harkrider[1], David Thornton[1], Megan Cuellar[3] and Tim Saltuklaroglu[1]*

[1] *Department of Audiology and Speech Pathology, University of Tennessee Health Science Center, Knoxville, TN, USA*
[2] *Department of Communication Disorders, University of Arkansas, Fayetteville, AR, USA*
[3] *Speech-Language Pathology Program, College of Health Sciences, Midwestern University, Chicago, IL, USA*

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Iiro P. Jääskeläinen, Aalto University, Finland
Anna J. Simmonds, Imperial College London, UK

****Correspondence:***
Tim Saltuklaroglu, Department of Audiology and Speech Pathology, University of Tennessee Health Sciences Center, 553 South Stadium Hall, UT, Knoxvilee TN 37996, USA
e-mail: tsaltukl@uthsc.edu

Activity in anterior sensorimotor regions is found in speech production and some perception tasks. Yet, how sensorimotor integration supports these functions is unclear due to a lack of data examining the timing of activity from these regions. Beta (~20 Hz) and alpha (~10 Hz) spectral power within the EEG μ rhythm are considered indices of motor and somatosensory activity, respectively. In the current study, perception conditions required discrimination (same/different) of syllables pairs (/ba/ and /da/) in quiet and noisy conditions. Production conditions required covert and overt syllable productions and overt word production. Independent component analysis was performed on EEG data obtained during these conditions to (1) identify clusters of μ components common to all conditions and (2) examine real-time event-related spectral perturbations (ERSP) within alpha and beta bands. 17 and 15 out of 20 participants produced left and right μ-components, respectively, localized to precentral gyri. Discrimination conditions were characterized by significant (pFDR < 0.05) early alpha event-related synchronization (ERS) prior to and during stimulus presentation and later alpha event-related desynchronization (ERD) following stimulus offset. Beta ERD began early and gained strength across time. Differences were found between quiet and noisy discrimination conditions. Both overt syllable and word productions yielded similar alpha/beta ERD that began prior to production and was strongest during muscle activity. Findings during covert production were weaker than during overt production. One explanation for these findings is that μ-beta ERD indexes early predictive coding (e.g., internal modeling) and/or overt and covert attentional/motor processes. μ-alpha ERS may index inhibitory input to the premotor cortex from sensory regions prior to and during discrimination, while μ-alpha ERD may index sensory feedback during speech rehearsal and production.

Keywords: speech perception, speech production, EEG, mu rhythm, independent component analysis

INTRODUCTION

It remains critical to disentangle the neural networks that allow an infinite array of co-articulated vocal tract gestures to be produced by a speaker and effortlessly sensed, recognized, and understood by a listener. Though these two complimentary and highly integrated processes often are examined independently, considerable recent effort has focused upon understanding how classical production mechanisms (e.g., the motor system) are involved in speech perception (D'Ausilio et al., 2012; Mottonen and Watkins, 2012; Murakami et al., 2013) and classical perception regions (i.e., auditory and somatosensory systems) are involved in production (Burnett et al., 1998; Stuart et al., 2002; Purcell and Munhall, 2006). Sensorimotor integration (SMI) provides an interface for speech perception and production and is fundamental to efficient verbal communication (e.g., Perrier et al., 1996; Rogalsky et al., 2011; Tourville and Guenther, 2011; Guenther and Vladusich, 2012; Moulin-Frier and Arbib, 2013). However, questions regarding the nature and timing of SMI prevail and relatively few studies address SMI in both speech perception and production within the same experiment (Wilson et al., 2004; Pickering and Garrod, 2007; Hickok et al., 2011; Adank, 2012).

Neuroimaging techniques have identified the auditory dorsal pathway (posterior temporal lobe, inferior parietal lobe, premotor cortex; PMC) as playing a role in both speech perception and production. In production, there is clear evidence of cooperation between feedforward and feedback systems for motor control (Houde and Nagarajan, 2011). Within speech perception, SMI is explained through independent yet convergent "dual streams" of neural activity (Scott and Johnsrude, 2003; Hickok and Poeppel, 2004; Hickok, 2009, 2012; Rauschecker, 2012; Specht, 2013). The ventral stream (predominantly within auditory regions) provides speech decoding and comprehension. The dorsal stream

(including activity from sensorimotor regions) is thought to provide an audio-motor interface linking auditory to articulatory goals in speech perception. Though dorsal stream activity has been reported to be left-lateralized, there is recent evidence of bilateral organization (Cogan et al., 2014; Simmonds et al., 2014).

Despite evidence that the motor system is relatively inactive in non-degraded passive listening conditions (Scott et al., 2009; Szenkovits et al., 2012) and lesions demonstrating that damage to motor regions has little effect on the ability to perceive speech (Rogalsky et al., 2011), a number of perception tasks have been identified in imaging studies in which motor regions are recruited. These conditions typically have been those in which task demands are increased and include categorical discrimination of foreign phonemes (Callan et al., 2006), phoneme segmentation (Burton et al., 2000; Locasto et al., 2004; Burton and Small, 2006), and speech in noise (Osnes et al., 2011; Alho et al., 2012; D'Ausilio et al., 2012). Thus, motor system activity in speech perception may be context dependent, in addition to being variable across individuals (Szenkovits et al., 2012).

Given equivocal findings, the role of the motor system in speech perception is hotly debated. Perhaps a more pertinent question is the extent to which dorsal stream motor activity functionally enhances the perceptual process (Gallese et al., 2011). Hickok et al. (2011) maintain that contributions of the motor system are strictly modulatory and depend on the cognitive demands associated with a particular task. Others support a more functional role (Binder et al., 2004; Meister et al., 2007; D'Ausilio et al., 2009; Sato et al., 2009; Osnes et al., 2011; Grabski et al., 2013; Mottonen et al., 2013). In these studies, dorsal stream articulatory-motor based speech representations are associated with accurate speech perception in some tasks. However, to bolster understanding of the functional contributions of dorsal stream motor activity in speech perception, it is necessary to address the time-course of activity relative to acoustic stimulation in addition to task performance.

For example, Callan et al. (2010) used a combination of fMRI and magnetoencephalography (MEG), measuring PMC activity in a forced-choice, syllable discrimination in noise task. For correct discriminations, activity in the PMC preceded and immediately followed acoustic stimulation. These findings were interpreted as PMC activity functionally aiding in speech perception and were explained from a Constructivist perspective. That is, previous sensorimotor experiences bestow the motor system with the capacity to provide early top-down influences (in the form of predictive internal models) to help constrain sensory analysis and aid in perception (Sohoglu et al., 2012). In a manner that is also consistent with earlier analysis-by-synthesis theories (Stevens and Halle, 1967), these data suggest motor activity should be maintained while the internal model (i.e., hypothesis) is compared to the sensory consequences. Had motor activity in this study been found in a different time frame, other explanations might arise. For example, if motor activity only coincided directly with the occurrence of acoustic stimuli and was not related to functional performance, it might be interpreted from a Direct Realist viewpoint (Fowler, 1986), as a motor reflection of sensory stimulation. Similarly, motor activity that followed acoustic offset by 200 ms or more might be interpreted as covert rehearsal while the acoustic stimuli are kept in working memory (Callan et al., 2010).

Oscillatory models offer a time-sensitive means of examining neural processing of speech. These models posit a strong relation among phases of delta, theta, and gamma oscillations, and the temporal envelope of speech with respect to the encoding of discrete speech units (e.g., syllables). This relation reflects further evidence of auditory-motor coupling grounded in evolutionary adaptation for efficiency (Ghitza et al., 2012; Giraud and Poeppel, 2012). Measuring changes in spectral power across beta (15–25 Hz) and alpha (8–13 Hz) frequency bands may offer an additional method for understanding sensorimotor processing. Beta suppression is often associated with the anticipation of performance (Gladwin et al., 2006; Arnal, 2012; Bickel et al., 2012; Zaepffel et al., 2013) of motor activity and predictive (i.e., a priori) top-down coding for sensory analysis. Alpha bands dominate the human brain and the enhancement or suppression of alpha band power often is considered an indicator of cortical activation/inhibition (Klimesch, 2012). Event-related alpha desynchronization (ERD) is considered a release from inhibition for sensory gating and may also contribute to predictive coding. In addition, alpha power generally is suppressed with increased attentional and cognitive demands. Weisz and colleagues provide evidence of an independent auditory alpha generator, implicating a link to speech perception (Weisz et al., 2011; Obleser and Weisz, 2012). Additionally, in support of alpha sensitivity to speech perception, they found that magnitude of alpha suppression across a broad (prefrontal, temporal, parietal) network corresponded with reductions in speech stimulus intelligibility.

The rolandic mu (μ) rhythm is characterized by an arc-shape, alpha and beta band peaks, and typically localized to sensorimotor regions (Pineda, 2005; Hari, 2006). Spectral power within the μ-rhythm is often considered a down-stream measure of motor activity from the PMC (Pineda, 2005). Suppression of the power in the alpha band of the μ-rhythm (μ-alpha) has been used to measure sensorimotor activity in response to viewing biologically relevant (i.e., reproducible) vs. non-relevant visual stimuli such as hand (Oberman et al., 2005; Perry and Bentin, 2010) and face (Muthukumaraswamy and Johnson, 2004) movements, visually presented speech (Crawcour et al., 2009), and motor imagery tasks (Tamura et al., 2012; Holler et al., 2013). μ-alpha also suppresses to action-based sounds (Pineda et al., 2013), speech stimuli in segmentation tasks, and when identifying speech in noise (Cuellar et al., 2012). Additionally, Tamura et al. (2012) reported μ-alpha suppression to overt and imagined speech production under various types of auditory feedback. Their findings suggest that μ-alpha suppression in speech provides an index of feedback in audio-vocal monitoring. This interpretation seems logical considering that μ-alpha suppression is thought to arise from somatosensory activity when guidance is needed for ongoing movement (Hari, 2006). Considering also that μ-beta suppression is indicative of motor activity, identifying patterns of μ-alpha and μ-beta ERS/ERD across speech tasks is likely to reveal further important information about the timing of motor and sensory contributions to SMI in speech processing.

To this end, Bowers et al. (2013) recently employed an EEG technique to study SMI during speech perception, adapting a

similar design from an fMRI and MEG study (Callan et al., 2010). Specifically, participants passively listened to and actively discriminated (i.e., forced choice, same or different) between pairs of syllables (/ba/ and /da/) and tone sweeps presented in different signal-to-noise ratios (SNRs). Raw data from 30 EEG recording channels were analyzed via independent component analysis (ICA). ICA is blind-source separation (i.e., linear decomposition) tool that can be used both as a strong filter and a means of independent and spatially fixed sources of neural activity (Delorme and Makeig, 2004; Makeig et al., 2004; Onton et al., 2006). Left and right μ-rhythm component clusters with characteristic spectral peaks at ∼10 Hz and ∼20 Hz (Hari, 2006), maximally localized to the sensorimotor cortex with activation extending into the PMC, were identified in most participants. Time-frequency analysis of μ components using event-related spectral perturbation (ERSP) analysis showed ERD in the beta band that was strongest when speech was accurately (>95% correct) discriminated in noise with a SNR of + 4 dB. Most importantly, in this condition only, μ-beta suppression (i.e., motor activity) began prior to speech perception and peaked immediately following stimulus offset. The findings were interpreted in accord with Callan et al. (2010) and others, suggesting that PMC/sensorimotor regions can readily contribute to speech perception (e.g., Skipper et al., 2007). From an oscillatory perspective, they were interpreted as evidence of early top-down influences from the motor system (i.e., internal models), helping to constrain auditory analysis in shared channels between sensorimotor and auditory regions (Arnal and Giraud, 2012).

Bowers et al. (2013) demonstrated that this event-related EEG technique with subsequent ICA/ERSP analysis is suitable for measuring SMI in speech perception. However, it is important to note that all their conditions employed background noise. Mottonen et al. (2013) used rTMS to impair motor representations from the lips and found impaired speech discrimination, suggesting that auditory-motor dorsal stream activity is important for speech discrimination in normal as well as degraded conditions. Alho et al. (2012) reported similar findings using evoked potentials. More generally, μ suppression has been found in anticipation of correctly predicted visual targets suggesting functional support in the task from attentional/motor networks (Bidet-Caulet et al., 2012). Hence, a salient question that remains pertains to the extent to which patterns of beta suppression in noisy speech discrimination tasks can disassociate the influences of a degraded listening environment from those functionally related to accurate categorical perception (Specht, 2014). To further understand how μ rhythms respond in speech discrimination tasks, it is necessary to examine the time-course of μ ERS/ERD in a quiet discrimination condition.

Regions within the dorsal stream help mediate sensorimotor control in speech production (Houde and Nagarajan, 2011; Hickok, 2012; Rauschecker, 2012). Feed-forward motor plans are generated that, once properly trained, allow for fluid generation of co-articulated speech gestures at an appropriate speaking rate (Tourville and Guenther, 2011). In addition, inverse forward models (i.e., efference copies) of predicted sensory consequences are sent from motor regions (i.e., premotor/motor cortex) to higher order auditory (e.g., superior temporal sulcus) and somatosensory (e.g., inferior parietal lobe) sites for dynamic comparisons with auditory and somatosensory production, providing the neurophysiological basis for an ongoing feedback loop. As forward prediction is compared with the intended targets and subsequently integrated with the true sensory (i.e., acoustic) and somatosensory consequences of production, corrective feedback is returned to motor control centers in a manner so efficient as to allow for online correction should a mismatch occur between predicted consequence and the articulatory goal. According to dynamic state feedback control (SFC) models, across the time course of any given speech production, complex dynamic interplay can exist between feedforward and feedback control in response to ongoing changes in vocal tract configurations and acoustic/somatosensory perturbations (Ventura et al., 2009; Houde and Nagarajan, 2011; Golfinopoulos et al., 2011; Hickok, 2012, 2014). Hence, as in speech perception, the addition of temporal data from regions within dorsal stream networks is likely to help foster a better understanding of the feedforward and feedback dynamics in speech production.

The largest obstacle to deploying imaging techniques with high temporal resolution such as EEG and MEG to speech production is signal contamination from muscle artifact. It is well known that myogenic activity from eyes (e.g., blinking), lips, head, and jaw produces robust electrical activity in frequency ranges broad enough to spuriously influence most neural activity. In addition, due to volume conduction, the effects of myogenic activity are not focal and can influence recordings from all cranial electrodes (McMenamin et al., 2011). Due to this limitation, EEG and MEG studies targeting language production networks have employed a variety of experimental designs intended to circumvent overt speech production. These designs have typically involved delayed or covert speech production. As evidence exists showing similarities in neural activity in overt and covert production tasks, Tian and Poeppel (2010, 2012) including the generation of internal models (Sams et al., 2005; Tian and Poeppel, 2010), covert production often provides a viable substitute for overt production tasks. However, in terms of SMI, the two tasks are different and may not share all the same neurophysiology (Ganushchak et al., 2011), especially in some pathological conditions with compromised sensorimotor control such as stuttering (Max et al., 2003; Loucks and De Nil, 2006; Watkins et al., 2008; Hickok et al., 2011; Cai et al., 2014; Connally et al., 2014). Other studies have stopped short of measuring activity during production, instead relying on oscillatory data from a time window prior to actual production. In this vein, a number of ERP studies have measured 'lexical' access and morphological encoding strategies (Hirschfeld et al., 2008; Costa et al., 2009; Dell'acqua et al., 2010; Strijkers et al., 2010). Whole head MEG data have revealed patterns of μ-alpha suppression in auditory regions with μ-beta suppression in auditory-motor (i.e., dorsal stream) integration regions (Gehrig et al., 2012). Similarly, Herman et al. (2013) measured real-time changes in oscillatory data from syllable encoding and pre-production time periods to identify discrete input/output operations within the dorsal stream phonological loop, again highlighting the value of temporal information.

Improvements in source estimations and data analysis techniques, along with continued widespread availability appear to

be contributing to a resurgence of EEG. ICA has been suggested as an effective technique for separating neural from myogenic activity on the basis of the assumption of temporal and spatial independence of components. Therefore, especially when stereotypical in nature, myogenic activity can be separated from neural activity in the unmixing process following ICA training on sufficient data (Delorme and Makeig, 2004; Onton et al., 2006; Gwin et al., 2010). The use of ICA in this capacity has been demonstrated to remove movement artifact while performing hand movements (Shou et al., 2012), walking and running (Gwin et al., 2010; Lau et al., 2014), and in distinguishing distinct patterns of electro-cortical activity in knee vs. ankle movements (Gwin and Ferris, 2011). However, its application to speech production has been limited. Though seemingly daunting, Tran et al. (2004) reported successfully using ICA to remove artifact from stuttered speech in children. In addition, other studies have demonstrated that ICA can be used to reveal neural activity not evident in univariate analyses (Geranmayeh et al., 2012; Simmonds et al., 2014).

Though there is reason for optimism regarding the potential use of EEG with ICA for measuring neural activity in speech production tasks (Ganushchak et al., 2011), concerns remain regarding the potential for ICA to adequately remove all muscle artifact. These include reduced validity for localization in intracerebral space, the fact that muscle artifact is often non-stereotypical and therefore not always suited for identification via ICA, and that a substantial portion of the variance in the whole EEG signal (i.e., up to 67% of components) can be accounted for by pure myogenic activity, reducing spectral power in neural components of interest (McMenamin et al., 2009; Shackman et al., 2009; McMenamin et al., 2010, 2011). It is clear that preliminary investigations using ICA in speech production should proceed cautiously using simple productions.

As a launching point, the current study focuses on activity from μ components for the following reasons. First, they are ubiquitously found in EEG recordings, particularly when identified via ICA. Thus, the possibility of yielding μ components in ICA decomposition remains high even when muscle components predominate. Second, μ rhythms typically are localized to primary motor/PMC regions, which are key sites within the dorsal stream. The PMC, in particular, is bi-directionally connected to higher-level auditory and somatosensory regions via the arcuate and longitudinal fasciculi. Its location and connectivity allow it to serve as an important intermediary for integrating forward prediction (internal modeling) and sensory feedback in both perception and production (Houde and Nagarajan, 2011; Rauschecker, 2012). Third, μ-ERS/ERD already has revealed real-time data interpreted as predictive coding in speech perception (Bowers et al., 2013). Alpha and slow beta bands which are contained within the μ rhythm are the only frequency domains that display ERS/ERD sensitivity to stimulus and/or task (Klimesch, 2012). Therefore, further time-frequency analyses of μ rhythms potentially may reveal important information about SMI in production.

There are two main goals in this study. The first is to bolster understanding of the timing and function of dorsal stream activity in speech perception by examining ERS/ERD patterns in quiet and noisy discrimination conditions. The second is to provide initial evidence that, via the application of ICA/ERSP, the use of EEG can be extended into the realm of speech production. Collectively, the intention is to show that ICA can be used accurately to identify dorsal stream sensorimotor μ components common to both speech perception and production. It is first hypothesized that right and left μ components, localized to sensorimotor/PMC regions, will be found across perception and production conditions. By placing EMG electrodes on the upper and lower lip, it also is anticipated that ICA will identify prominent perilabial muscular activity. Once μ components are identified, the second hypothesis is that ERSP analyses will provide differential time-frequency measures of alpha and beta ERS/ERD. Real-time oscillatory changes in the spectral power of alpha and beta bands of the μ rhythm are expected to provide novel information regarding the timing of SMI in speech perception and production that may be interpreted via dual stream/SFC models. Additionally, significant activity from perilabial components is expected only in overt production, allowing it to be mapped in real-time to SMI activity.

MATERIALS AND METHODS

PARTICIPANTS

Twenty right-handed English-speaking adults (17 female and 3 males) with a mean age of 23.94 years (range 21–39 years) were recruited from audiology and speech pathology classes at the University of Tennessee Health Science Center. Participants reported no diagnosed history of communicative, cognitive, or attentional disorders. Handedness dominance was assessed using the Edinburgh Handedness Inventory (Oldfield, 1971). This study was approved by the Institutional Review Board of the University of Tennessee Health Science Center. Prior to the experiment, participants provided signed informed consent on a document approved by the Institutional Review Board.

STIMULI

Perception

/ba/ and /da/ syllables were created using AT&T naturally speaking text-to-speech software which employs synthetic analogs of a human male speaker. Syllable pairs were generated such that half of the stimuli were composed of different syllables (e.g., /ba/ and /da/) and the other half were identical (e.g., /ba/ and /ba/). The stimuli were low-pass filtered below 5 KHz and normalized for root-mean-square (RMS) amplitude. Each syllable was 200 ms in duration. Each syllable pair was also separated by 200 ms, resulting in a total of 600 ms from the first syllable onset to the second syllable offset (**Figure 1**).

For one condition (discrimination in noise; Ndis), syllable pairs to be discriminated were embedded in white noise with a SNR of +4 dB. This SNR was chosen as it has been shown previously (Bowers et al., 2013) to produce discrimination accuracies > 95% in a similar group of participants. In the other discrimination condition (discrimination in quiet; Qdis), syllable pairs were presented without background noise. To prevent discrimination response bias (Venezia et al., 2012), in both Qdis and Ndis stimuli sets, there were equal numbers of syllable pairs that were identical as there were different.

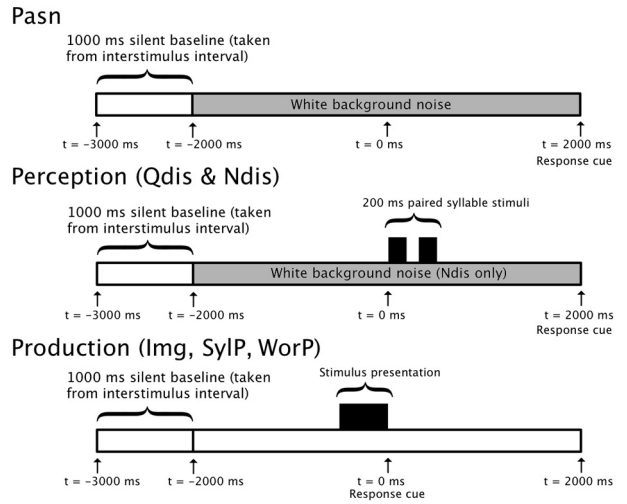


FIGURE 1 | Single-trial timelines for perception and production conditions. Timelines for 5000 ms (−3000 to 2000 ms) in passive noise condition (Pasn), discrimination conditions (Qdis & Ndis), and production conditions (Img, SylP, and WorP).

Production

Speech targets were syllable pairs, similar to those used in the discrimination tasks above, and tri-syllable nouns (initiated with /b/ or /d/ and followed by a vowel). They were displayed centered in Microsoft PowerPoint slides with plain black backgrounds in large white Arial font (size 56). **Figure 1** shows the timelines for epochs in both perception and production conditions.

DESIGN

A 6-condition within-subject design was employed. Based on extant literature, the conditions were created to increase motoric demands incrementally (i.e., from the perceiving of white noise to overtly producing tri-syllable words). Participants were required to:

(1) passively listen to white noise (Pasn).
(2) discriminate (same or different) between pairs of syllables in a quiet background (Qdis).
(3) discriminate (same or different) between pairs of syllables in a noisy background (Ndis).
(4) imagine producing a pair of syllables (Img).
(5) overtly produce (i.e., say) a pair of syllables (SylP).
(6) overtly produce (i.e., say) tri-syllable nouns initiated by /b/ or /d/ and followed by a vowel (WorP).

Thus, condition 1 (Pasn) required no discrimination and was a control task for the two discrimination conditions (Qdis and Ndis). Conditions 2–5 employed /ba/ and /da/ syllables. Conditions 2 and 3 (Qdis and Ndis) required same/different discriminations of random /ba/ and /da/ combinations, while conditions 4 and 5 required covert (Img) and overt (SylP) production of randomly selected /ba/ and /da/ combinations. Condition 6 (WorP) also required overt production, but in this condition tri-syllable nouns were used as opposed to the 2-syllable combinations employed in the SylP condition. In the WorP condition, words meeting these criteria were selected from Blockcolsky et al. (2008). Examples of these words include "dialog," "butterscotch," "daffodil," and "buffalo."

PROCEDURE

The experiment was conducted in an electronically and magnetically shielded, double-walled, sound-treated booth. Participants sat in a comfortable reclining armchair with their heads and necks well supported. Compumedics NeuroScan Stim 2 version 4.3.3 software was used to present stimuli to participants via a PC computer and record button-press responses. A button-press response was required for all three perception conditions because anticipation of a button-press has previously been known to elicit μ-rhythm ERD (Makeig et al., 2004; Graimann and Pfurtscheller, 2006; Hari, 2006). Hence, in the Pasn condition, the button-press was used as a control for the required button-press response in the discrimination conditions and to ensure that participants were paying attention in each trial. The cue to respond was a 100 ms, 1000 Hz tone that was presented at the end of the epoch (i.e., +2000 ms). In the Pasn condition, participants were instructed simply to listen passively to the noise and press the designated button after hearing a pure tone cue in each trial. Designation of button-press responses (right or left hand) was counterbalanced across all subjects and experimental conditions. Performance in the discrimination conditions was evaluated by calculating the percentage of correct trials.

In the production conditions, stimuli appeared on a 69.5 × 39.0 cm display placed 132 cm in front of the reclining chair. The stimuli appeared on the screen for 1 s. Participants were instructed to begin their production response immediately when the stimulus disappeared from the monitor. In the Img condition participants were told to imagine saying (i.e., covertly producing) the pair of syllables while refraining from making any overt articulatory movements or vocalization. In the SylP and WorP condition, participants were instructed to speak the syllable pair or word in their normal speaking voice. All overt speech productions were easily completed in the time window (2 s) following the cue to speak. All conditions were presented in two blocks of 40 trials each. The order of the 12 blocks (6 conditions × 2 blocks) was randomized for each participant.

EEG ACQUISITION

Sixty-eight electrode channels were used to acquire whole-head EEG data. These included two electromyography (EMG) and two electrocardiogram (ECG) electrodes. Electrode configuration was based upon the extended international standard 10–20 (Jasper, 1958) method using an unlinked, sintered NeuroScan Quik Cap (Towle et al., 1993). All recording electrodes were referenced to the common linked left (M1) and right (M2) mastoids. The electro-oculogram (EOG) was recorded by placing electrodes on the left superior orbit and the left inferior orbit (VEOG) as well as the lateral and medial canthi of the left eye (HEOG) to monitor vertical and horizontal eye movements, respectively. The two surface electromyography (EMG) electrodes were placed at midline above the upper lip and below the lower lip for the

purposes of collecting perilabial EMG data related to speech production.

EEG data were collected using Compumedics NeuroScan Scan 4.3.3 software and the Synamps 2 system. The raw EEG data were filtered (0.15–100 Hz) and digitized via a 24-bit analog-to-digital converter at a sampling rate of 500 Hz. Data collection was time-locked to time point zero at the onset of acoustic stimuli delivery in speech perception trials and the cue to begin speaking in production trials. Thus, in the perception conditions, time zero referenced the acoustic onset of the first syllable. In the production conditions, syllable and word stimuli were orthographically displayed on the monitor between times -1000 ms and zero. Hence, disappearance of the text at time zero was the cue for participants to begin speaking (**Figure 1**).

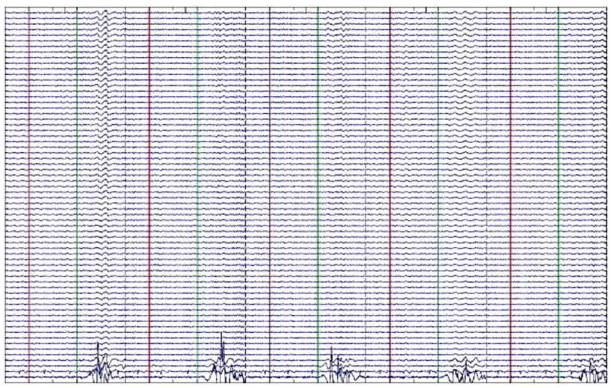


FIGURE 2 | Stereotypic muscle activity from perilabial EMG electrodes. Example of filtered and epoched EEG data from the WorP condition showing stereotypical EMG activity during speech production.

EEG DATA PROCESSING

EEGLAB 12 open source software (Delorme and Makeig, 2004) was used to process all EEG data by performing the following steps for individual and group processing/analysis.

(1) Individual processing/analysis:

  (a) 12 raw EEG files (6 conditions x 2 blocks) were pre-processed for each participant.
  (b) Independent component analysis (ICA) was performed on all concatenated files across all conditions for each participant.
  (c) All neural and non-neural dipoles were localized for each independent component (IC) identified.

(2) Group analysis:

  (a) Using the STUDY module of EEGLAB 12, two separate analyses were performed using 'in head' only (neural) and 'all' (neural and non-neural) ICs.
  (b) Principal component analysis (PCA) subsequently was used to identify and cluster common components across participants.
  (c) Left and right μ clusters were identified from the 'in-head' STUDY, whereas the EMG cluster representing perilabial muscle activity was identified from the 'all' STUDY.
  (d) μ clusters were localized using equivalent current dipole (ECD) and current source density (CSD) analyses.
  (e) Time-frequency analyses (measuring changes in spectral power across time) were performed by measuring event-related spectral perturbations (ERSP) in the left and right μ clusters as well as in the EMG cluster.

Details of each step in the data processing / analyses are described below.

Processing/analysis of EEG data from each participant

Data pre-processing. Raw data from both 40-trial blocks in each condition were: (1) appended to make a single 80 trial data set for each condition; (2) downsampled to 256 Hz to decrease computational requirements for ICA processing; (3) epoched into 5000 ms segments with individual epochs spanning from −3000 to +2000 ms around time zero; (4) band-pass filtered (3–34 Hz) to ensure that alpha and beta could be identified while filtering muscle movement from surrounding frequencies; (5) re-referenced to mastoid electrodes; (6) visually inspected for gross artifact ($> 200\,\mu V$), which was manually removed; and (7) pruned to remove trials with incorrect responses or response latencies greater than 2 s in the Qdis and Ndis conditions, although few trials were removed (see below). A minimum contribution of 40 epochs per participant per condition was required for inclusion in the experiment. However, the average number of usable trials across participants per condition far exceeded the minimum of 40 required for inclusion. **Figure 2** shows an example of EEG activity from 5 trials in one participant in the WorP condition following filtering and epoching. Critically, the muscle activity from the EMG component appears to be relatively stereotypical in nature (e.g., **Figure 2**), thereby facilitating ICA efforts to separate the neural activity from the muscle activity in the subsequent ICA signal decomposition.

Independent component analysis (ICA). Prior to ICA training, pre-processed EEG data for each participant were concatenated across all 6 conditions so that a single set of ICA weights could be obtained. This allowed for a comparison of activity to be made across conditons within spatially fixed ICs. An extended Infomax algorithm (Lee et al., 1999) was used to decorrelate the data matrix prior to ICA rotation. ICA training was provided using the "extended *runica*" algorithm in EEGLAB 12 with an initial learning rate set to 0.001 and a stopping weight of 10–7. Following decomposition, 66 ICs were yielded for each participant reflecting the total number of recording electrodes (68 – 2 reference electrodes, M1 and M2). Scalp maps for each IC were obtained by projecting the inverse weight matrix (W^{-1}) back onto the spatial EEG channel configuration.

Following ICA decomposition, equivalent current dipole (ECD) models for each component were computed using a boundary element model (BEM) in the DIPFIT toolbox, freely

available at sccn.ucsd.edu/eeglab/dipfit.html (Oostenveld and Oostendorp, 2002). Standard 10–20 electrode coordinates were warped to the head model followed by automated coarse-fitting to a BEM, yielding a single dipole model for each of 1320 ICs (66 ICs × 20 participants). Dipole localization requires back-projecting the signal to a source that may have generated the scalp potential distribution for a given IC, and then computing the best forward model to explain the highest percentage of scalp map variance (Delorme et al., 2012). Residual variance (RV) in dipole localizations were also computed, referring to the potential mismatch between the initial scalp map and the forward projection of the ECD model.

Group data analyses

EEGLAB STUDYs. Group data analyses were conducted via the EEGLAB STUDY module. The STUDY module allows ICA data from multiple participants across conditions to be analyzed using specified designs. In the current study, the designs specified were dictated by the within-subjects conditional differences of interest. The STUDY module allows further filtering to be applied with respect to the RV in dipole localization and inclusion vs. exclusion of out-of head dipoles. Thus, ICA files with dipole information from each individual (see above) were applied to the two separate STUDY modules. For the purposes of measuring neural activity, only "in-head" dipoles with RV< 20% were analyzed.

For the purposes of identifying perilabial EMG activity, a second STUDY was conducted that included "all" dipoles from in head and outside the head. In this second STUDY, the RV criterion was raised to 50% (Gramann et al., 2010) dipoles because EMG activity emanates from outside the head and by nature, muscular movement incurs higher unexplained RV.

Principal component clustering of ICs. In both the "in head" and "all" STUDYs, component pre-clustering was performed on the basis of common scalp maps, dipoles, and spectra. The K-means statistical toolbox (implemented in EEGLAB; Delorme and Makeig, 2004) then used these criteria to group similar components from each participant via PCA. After removal of outliers (3 SD from any cluster mean), components from the "in head" STUDY were assigned to 20 possible neural clusters, which included left and right sensorimotor μ clusters. Components in the "all" STUDY were assigned to 66 possible clusters and included one non-neural cluster depicting perilabial EMG activity.

Final component designation to left and right μ clusters was based primarily on the PCA followed by individual inspection of spectra, scalp maps, and dipoles of all components within those clusters and neighboring clusters. Final inclusion criteria for membership to μ clusters included localization to BA 1–4, and 6 (i.e., somatosensory regions, primary motor and premotor regions) and characteristic μ spectra, though over 90% of components emanated from BA 6.

Components in the "all" STUDY were assigned to 66 possible clusters, most of which, as expected, depicted non-neural activity. The cluster characterizing perilabial EMG activity was found on the basis of dipole location and ERSP analysis showing activity only in overt production tasks (see below).

μ cluster source localization. ECD source localization is simply from the average (x, y, z) coordinate of all the IC dipoles (identified via the DIPFIT module) within a given cluster. Alternatively, standardized low-resolution brain electromagnetic tomography (sLORETA) uses current source density (CSD) distribution from electrical potential measures across the scalp to address the inverse problem and provide an estimate of source localization (Pascual-Marqui, 2002). The head model uses a Talairach cortical probability brain atlas, digitized at the Montreal Neurological Institute (MNI). EEG electrode locations are cross-registered between spherical and realistic head geometry (Towle et al., 1993). Spatial resolution of 5 mm is achieved by sampling 6239 voxels in 3-D brain space. For each IC that contributed to the two μ clusters, the inverse weight projections on the original EEG channels were exported to the sLORETA. Cross-spectra were computed and mapped to the standard Taliarach brain atlas cross-registered with the Montreal Neurological Institute (MNI) coordinates, yielding sLORETA estimates of CSD for left and right μ dipoles in the "in-head" STUDY. To evaluate the statistical significance of dipole locations across participants, statistical comparisons relative to zero (i.e., no activation) were computed (Grin-Yatsenko et al., 2010). Paired (Student) *t*-tests were conducted on frequencies between 4 and 33 Hz (1000 frames) with the smoothing parameter set to 1 (single common variance for all variables), using 5000 random permutations yielding corrected *t*-value thresholds and statistical significance ($p < 0.001$) for all 6239 voxels.

While these two methods of EEG source localization were expected to produce similar results (Bowers et al., 2013), for reliability purposes it was deemed useful to use both techniques.

Time-frequency analysis (change in spectral power across time). ERSP analyses were used to compute changes (scaled in normalized dB units) in power across time (i.e., time-frequency analysis) within the spectral range of interest (4–33 Hz). Time-frequency transforms were derived using a Morlet sinusoidal wavelet set at 3 cycles at 3 Hz, rising linearly to 20 cycles at 40 Hz. The 1000 ms pre-stimulus period was selected from the silent inter-trial interval to serve as a baseline for each trial. These baselines were constructed from a surrogate distribution based on estimates of spectral power from 200 randomly selected latency windows from within the 1000 ms inter-trial interval (Makeig et al., 2004). Subsequent individual ERSP changes from baseline over time were computed using a bootstrap resampling method ($p < 0.05$ uncorrected). The single trial current for all experimental conditions for frequencies between 7 and 30 Hz and times from −500 to 1500 ms were entered in the time-frequency analyses.

In the "in-head" STUDY, differences in cross-conditional ERSPs in right and left μ clusters were computed using permutation statistics (2000 permutations) with a 95% confidence interval ($p < 0.05$). The random distribution represents the null hypothesis that no condition differences exist. Type I error was controlled by correcting conservatively for false discovery

rates (*p*FDR; Benjamini and Hochberg, 2000). Statistical analysis in the perception conditions used a 1 × 3 (Pasn, Qdis, Ndis) repeated measures ANOVA design. *Post-hoc* comparisons examined differences between Pasn vs. Qdis and Pasn vs. Ndis conditions. In the production conditions, a 1 × 3 repeated measure ANOVA design examined differences in ERSP activity across the Img, SylP, and WorP conditions. A *post-hoc* paired comparison examined differences between SylP and WorP conditions. In the "all" STUDY, cross-conditional ERSPs were computed in the production conditions using a 1 × 3 repeated measure ANOVA design.

RESULTS

DISCRIMINATION ACCURACY

In participants that contributed to μ clusters, the average number of useable trials (out of 80) across participants in each condition were: Pasn = 73.8 (*SD* = 7.2); Qdis = 74.8 (*SD* = 4.6); Ndis = 69.0 (*SD* = 11.4); Img = 75.0 (*SD* = 5.8); SylP = 71.1 (*SD* = 7.4); WorP = 69.9 (*SD* = 8.0). In the Qdis condition, all participants discriminated with 91–100% accuracy. In the Ndis condition, all except one participant discriminated with 84–100% accuracy. The remaining participant discriminated with 65% accuracy. The average discrimination accuracies in the Qdis and Ndis conditions were 97.3 and 94.4%, respectively. A paired *t*-test indicated that mean discrimination performance was not significantly different ($p > 0.05$) in these conditions. The average response latencies in the Qdis and Ndis conditions were 504 and 545 ms, respectively. A paired *t*-test again indicated that these latencies were not significantly different ($p > 0.05$). Together, these findings suggest that both discrimination tasks were performed with similar high levels of accuracy and efficiency. It should be noted again that trials with incorrect discriminations were eliminated from the data so the EEG analysis was limited to correct productions only.

μ AND EMG CLUSTER CHARACTERISTICS

As predicted by the first hypothesis, 17/20 and 15/20 participants produced components with< 20% unexplained RV that contributed to left and right μ clusters, respectively. For the left μ cluster, the average Talaraich ECD location was [−41, 4, 46], while on the right it was [46, 0, 39]. The percentage of unexplained RV in these single dipole models was 10.1 and 8.7% for the left and right hemispheres, respectively. sLORETA analyses revealed significantly activated voxels ($p < 0.001$) associated with μ clusters. Maximum current source densities were found at Talairach [−45, −10, 45] on the left vs. [45, −5, 40] on the right. In accord with findings by Bowers et al. (2013), the two localization techniques produced similar results, here allowing sources of μ activity to be maximally localized within the precentral gyri with activity spreading across the PMC and sensorimotor regions. **Figures 3** and **4** respectively display the scalp maps (A), spectra (B), ECD dipole clusters (C) and CSD maxima (D) for left and right μ clusters, respectively. The EMG cluster was characterized by non-neural ICs with an average of 21.3% unexplained RV.

TIME-FREQUENCY ANALYSES IN PERCEPTION (Pasn, Qdis AND, Ndis) CONDITIONS

Figure 5 shows Van Essen maps (generated using sLORETA) of significant voxels contributing to left (A) and right (B) μ clusters, followed by time-frequency (ERSP) analyses within

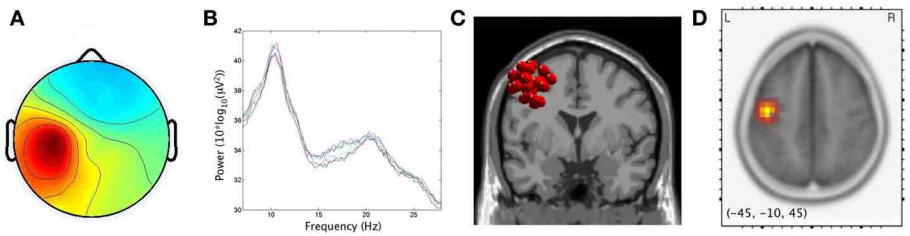


FIGURE 3 | Cluster results for left μ component. (A) mean scalp potential distribution (W^{-1}) scaled to RMS microvolts, **(B)** mean spectra of the components within the cluster for each condition, **(C)** distribution of equivalent current dipoles within the cluster, and **(D)** maximum current source density voxels (***t***-values) with greater values in darker colors and smaller values in lighter colors (at $p < 0.001$ corrected for multiple comparisons).

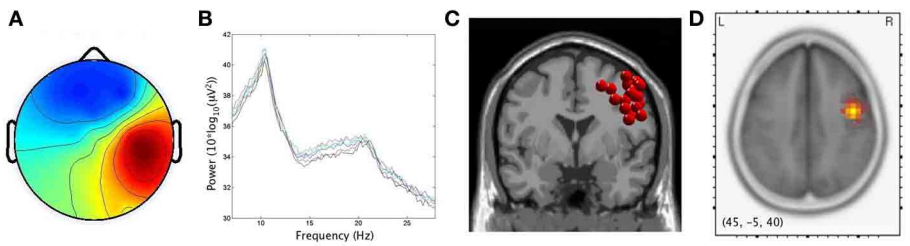


FIGURE 4 | Cluster results for right μ component. (A) mean scalp potential distribution (W^{-1}) scaled to RMS microvolts, **(B)** mean spectra of the components within the cluster for each condition, **(C)** distribution of equivalent current dipoles within the cluster, and **(D)** maximum current source density voxels (***t***-values) with greater values in darker colors and smaller values in lighter colors (at $p < 0.001$ corrected for multiple comparisons).

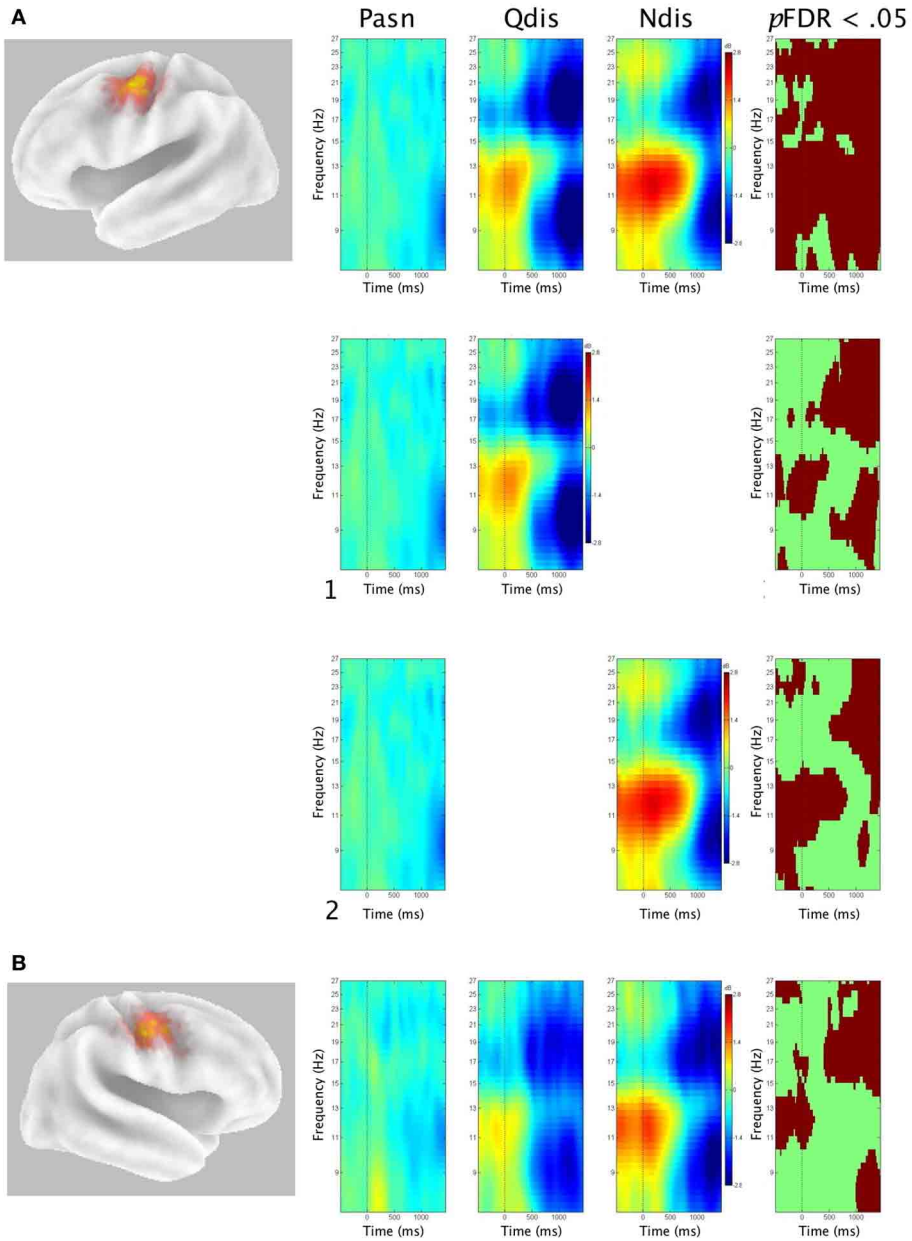


FIGURE 5 | Mean left and right ERSPs and sLORETA solutions for perception conditions. Rows A and B show sLORETA solutions for left and right μ clusters, respectively, depicted on a 3D Van Essen average template, followed by mean time-frequency ERSPs (event-related spectral perturbations) as a function of perception conditions, before, during, and after stimulus offset for **(A)** left μ clusters with (1) contrasts between Pasn and Qdis and (2) contrasts between Pasn and Ndis; and **(B)** right μ clusters (red, ERS, blue, ERD). The last frame in each row shows significant differences across conditions (pFDR< 0.05).

the 7–30 Hz bandwidth. The ERSP analyses show significant ERS/ERD changes from baseline in the Pasn, Qdis, and Ndis conditions. The last frame in each row shows statistical ERSP differences across conditions (pFDR < 0.05), thus supporting the second hypothesis.

For the left μ cluster, relative to the Pasn, alpha ERS began prior to acoustic stimulation and gradually gave way to alpha ERD beginning in low alpha frequencies (8–11 Hz) following acoustic offset in both discrimination conditions (Qdis and Ndis). Beta ERD in both discrimination conditions began in a narrow bandwidth (17–19 Hz), growing stronger and spreading across beta frequencies during and immediately following the acoustic stimulation condition. *Post-hoc* analyses (shown in **Figures 5A1,A2**) show differential patterns of significant beta ERD and alpha ERS/ERD in Pasn vs. Qdis comparisons and Pasn vs. Ndis comparisons.

Patterns of alpha/beta ERS/ERD were similar yet weaker and more diffuse in the right μ cluster compared to those on the left.

It followed that *post-hoc* ERSP comparisons of Qdis and Ndis to Pasn comparisons for right μ activity did not yield additional data of interest.

TIME-FREQUENCY ANALYSES IN PRODUCTION (Img, SylP AND, WorP) CONDITIONS

Figure 6 shows Van Essen maps (generated using sLORETA) of significant voxels contributing to left (A) and right (C) μ clusters, followed by time-frequency (ERSP) analyses within the 7–30 Hz bandwidth. The ERSP analyses show significant ERS/ERD changes from baseline in the Pasn, Qdis, and Ndis conditions. The last frame in each row shows statistical ERSP differences across conditions (pFDR < 0.05), again supporting the second hypothesis. **Figure 6B** shows the average ECD dipole location for the EMG components followed by ERSP analyses with statistical differences across conditions.

Significant EMG ERS (i.e., activity indicative of lip movement) in the SylP and WorP conditions began ∼300 ms after the cue to initiate speech. In both left and right μ clusters, alpha/beta ERD relative to baseline began in all production conditions up to 500 ms before the cue to speak. However, alpha/beta ERD in SylP and WorP conditions was significantly stronger (pFDR < 0.05) than in the Img condition during overt speech production (i.e., coinciding with EMG activity). *Post-hoc* analyses in both left and right μ clusters showed no ERSP differences in SylP vs. WorP conditions.

As μ-ERD was significantly weaker in Img relative to overt production (SylP and WorP) conditions, ERSPs for all components contributing to left and right μ clusters were examined in the Img condition. On the left, only 8 of 17 participants displayed μ-ERD in this condition. The others either showed ERS or negligible change. On the right, 6 of 15 showed μ-ERD, 2 showed patterns of alpha ERS with beta ERD, and the others showed either ERS or negligible change.

DISCUSSION

In accord with the aims and hypotheses of this study, left and right μ components were identified across perception and production tasks. 85 and 75% of participants submitted components with ∼10% unexplained RV in left and right μ components, respectively. This proportion of useable μ components is similar to that found in other studies (e.g., Nystrom, 2008; Bowers et al., 2013), though the proportion of unexplained RV is slightly higher, possibly due to the inclusion of motor tasks. Bilateral localization of μ rhythm source maxima to the precentral gyrus with activity spreading across the premotor and

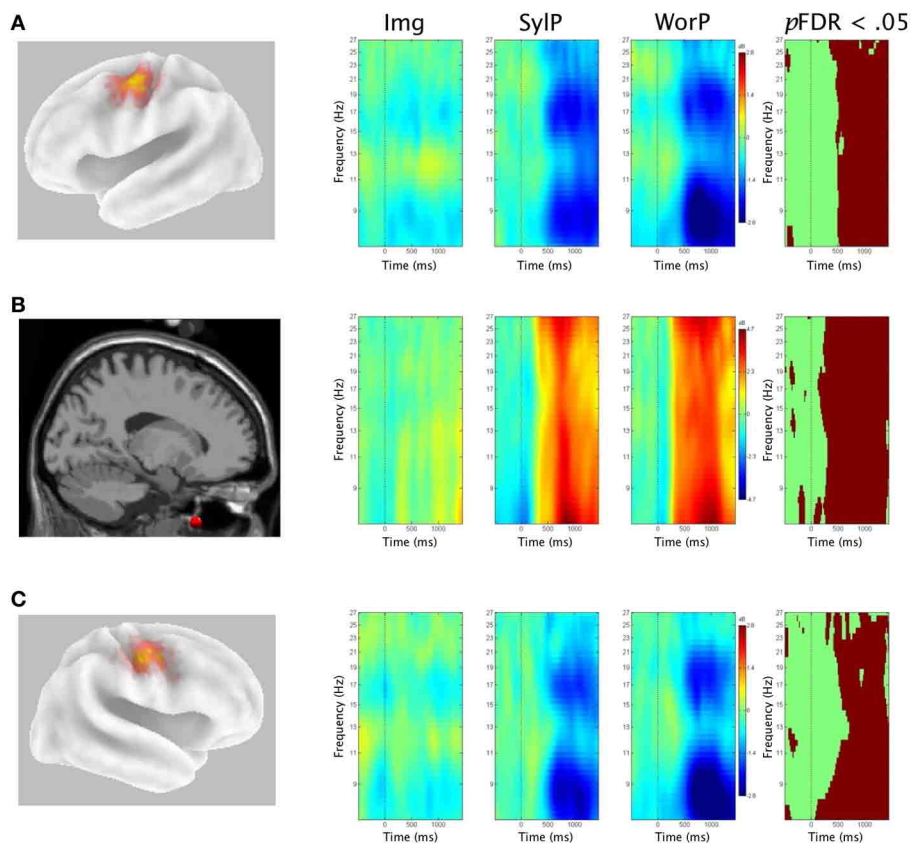


FIGURE 6 | Mean left and right ERSPs and sLORETA solutions for production conditions. Rows A and C show sLORETA solutions for left and right μ clusters, respectively, depicted on a 3D Van Essen average template, followed by mean time-frequency ERSPs (event-related spectral perturbations) as a function of production conditions, before, during, and after stimulus offset. The last frame in each row shows significant differences across conditions (pFDR < 0.05). Row B shows activity within the EMG component from perilabial myogenic activity.

sensorimotor cortices is consistent with accepted sources of the rhythm (Pineda, 2005; Hari, 2006) and important roles in speech perception/production (Skipper et al., 2007; Sato et al., 2009; Callan et al., 2010; Houde and Nagarajan, 2011; Tourville and Guenther, 2011). As these cortical sites are known to play important roles in SMI for both speech perception and production and μ rhythms are comprised of frequency bands that are sensitive to the demands of speech processing, this finding supports the subsequent examination of real-time activity within these clusters for better understanding the temporal dynamics of activity in the dorsal speech stream.

TIME-FREQUENCY ANALYSES IN PERCEPTION CONDITIONS

Similar to previous investigations (Makeig et al., 2004; Graimann and Pfurtscheller, 2006; Hari, 2006), anticipation of a button-press response in the Pasn condition yielded low-level increases from baseline in bilateral μ-ERD that gained slightly in strength with temporal proximity to the response. Discrimination conditions herein (Qdis and Ndis) also employed button-press responses, hence controlling for this effect in the statistical analysis. The Qdis and Ndis conditions produced similar highly accurate syllable discriminations and response reaction latencies such that differences between μ ERS/ERD in these conditions are attributable to the presence or absence of noise. Both conditions produced similar bilateral patterns of μ ERS/ERD that were generally stronger in the left hemisphere than the right, supporting left hemisphere dominance for SMI in speech perception (Hickok et al., 2011).

μ-alpha

Activity in the alpha band was characterized initially by ERS occurring prior to stimulus onset. ERS gradually gave way to ERD (**Figure 5**), with suppression first in low alpha (i.e., 8–10 Hz) and then high alpha (11–13 Hz). Alpha ERS was stronger and the transition occurred later in the Ndis condition than in the Qdis condition. Alpha rhythms are found globally across the cortex and their power can vary with numerous cognitive states and processes (Klimesch, 2012). Therefore, it is necessary to interpret alpha ERS/ERD relative to the tasks that induced them. Enhanced alpha (i.e., ERS) often is associated with cognitive load in working memory and attention tasks (Leiberg et al., 2006; Jensen et al., 2007; Haegens et al., 2010). It is thought to be an index of cortical inhibition of sensory information irrelevant to a given task, functioning to help sharpen attention to relevant information (Klimesch, 2012; Wilsch et al., 2014). In speech perception, this type of "active sensing" has been described in phenomena such as in the "cocktail party" effect, where specific attention to relevant speech cues helps filter similar competing background speech (Schroeder et al., 2010; Zion Golumbic et al., 2012).

Weisz et al. (2011) provide compelling evidence for an independently generated auditory alpha that is responsive to speech perception. Parsimonious with notions of increased cognitive load and consistent with the current findings, signal degradation of speech by noise vocoding has also been shown to enhance alpha activity (Obleser and Weisz, 2012). The observed differences in early alpha ERS between the Qdis and Ndis conditions support these notions. On the other hand, in speech perception tasks, alpha ERD has been found while evaluating speech (Shahin et al., 2009). Late occurring posterior alpha ERD has been related to increased speech intelligibility (Obleser and Weisz, 2012). Both of these findings are consistent with the notion of alpha ERD during accurate performance of perceptual and memory tasks (Klimesch et al., 2006). In addition, μ-alpha is suppressed in auditory speech perception tasks (Cuellar et al., 2012; Pineda et al., 2013). Hence, the current findings of late alpha ERD suggest that following stimulus offset, the two syllables were being evaluated by participants in the decision-making process.

μ-beta

In both discrimination conditions, significant beta ERD (relative to Pasn) was found across the time course of trials, prior to, during, and after acoustic stimulation. Beta ERD spread from narrow (17–19 Hz) to wide (15–30 Hz) beta bands while gaining in strength in both discrimination conditions. However, beta ERD occurred earlier in the Qdis than the Ndis condition. Bowers et al. (2013) previously showed that early beta ERD occurred when discriminating speech but not tones. They suggested that in speech perception tasks, early beta ERD also can be explained as a function of predictive coding. That is, internal models are posited to be generated in motor regions that are delivered to higher order auditory regions (i.e., superior temporal sulcus) to help constrain analysis and functionally improve speech discrimination accuracy (Callan et al., 2010; Bowers et al., 2013). These findings are also consistent with those of Mottonen et al. (2013), in that degraded conditions do no appear necessary to induce motor activity in speech discrimination. These models are thought to be available because of the considerable experience of humans generating the movements that produce these sounds. In addition, this predictive coding may have been fine tuned within the experiment. That is, requiring participants to accurately discriminate syllables 160 times (80 per condition) may have elicited anticipatory attention to speech processing.

μ-alpha and beta in discrimination conditions

The patterns of alpha and beta μ ERS/ERD found in quiet and noisy accurate speech discrimination need to be considered in combination. While similar patterns were observed in Qdis and Ndis conditions, stronger early alpha ERS was observed in the Ndis condition, which is consistent with the requirement of discriminating in noise. That is, it is speculated that the inhibitory mechanism was stronger when background noise was present. Conversely, early beta suppression appeared to be stronger in the Qdis than the Ndis condition. Though it is likely that internal models were generated in both conditions since they both used speech and were discriminated accurately (Bowers et al., 2013), it appears that in this study, the strong alpha ERS may have dominated the μ rhythm, extended into the low beta frequencies in the Ndis condition, and perhaps negated some early beta ERD. Together, these data suggest that alpha ERS and beta ERD within the sensorimotor μ rhythms work in unison, co-operating to functionally support accurate speech discrimination. This is further evidence that examination of the μ-rhythm provides a rich, time-sensitive, and relatively unique view of SMI in speech discrimination from an oscillatory perspective.

Dorsal stream motor activity in speech perception

The source location of μ clusters and their alpha and beta ERS/ERD suggest that they provide important information regarding sensorimotor dorsal stream activity in speech perception. The current findings suggest that beta activity may provide a measure of predictive coding via internal models (Bowers et al., 2013) generated in the PMC (Skipper et al., 2007; Houde and Nagarajan, 2011; Tourville and Guenther, 2011; Rauschecker, 2012). Tamura et al. (2012) investigated μ rhythm activity in various speech tasks, including covert production and production under different types of auditory feedback. They found differential activity within the alpha band and concluded that the μ-alpha was an index of auditory monitoring for speech. In line with this notion, it is speculated that μ-alpha might index sensory feedback into the PMC. Thus, stronger alpha ERS in the Ndis condition was observed, possibly due to a stronger inhibition of auditory feedback to the PMC when speech was presented in background noise. Furthermore, in the time period following stimulus offset and prior to the button press response, it seems likely that the two syllables were held in working memory, while being compared and covertly replayed during the decision-making process. These processes may require the generation of internal speech models and the disinhibition of feedback to the PMC, which would support the current findings of alpha and beta ERD in this time period within trials.

TIME-FREQUENCY ANALYSES IN PRODUCTION CONDITIONS

The covert (Img) and overt (SylP and WorP) production conditions yielded similar general patterns of alpha/beta ERD relative to baseline across trials. However, both alpha and beta ERD were significantly stronger in the overt production conditions than the Img condition, with significant differences in ERD following the cue to speak in both conditions. Across production conditions, there appeared to be little difference between right and left μ ERD. This is consistent with others that have found movement-induced bilateral decreases in beta suppression across the sensorimotor cortex (e.g., Salmelin and Hari, 1994; Pfurtscheller et al., 1996; Stančák and Pfurtscheller, 1996; Leocani et al., 1997, 2001; Alegre et al., 2003; Rau et al., 2003; Bai et al., 2005; Doyle et al., 2005; Erbil and Ungan, 2007). No differences between SylP and WorP conditions were observed. As expected, ERSP time-frequency analysis of perilabial EMG activity showed little activity in the Img condition, confirming that participants did not articulate the target syllables. In the SylP and WorP conditions, EMG activity following the "go" cue to speak was characterized predominantly by strong ERS beginning ∼300 ms in both conditions. This time lag from the "go" cue is consistent with a normal movement reaction time. Hence, μ-alpha and μ-beta ERD showed temporal alignment to lip muscle movements.

μ-alpha during production

μ-alpha ERD in speech production is again interpreted as an index of feedback to the PMC while speech is being produced. By only measuring activity in the sensorimotor μ, it is not possible to differentiate between auditory and somatosensory feedback. μ-alpha suppression is traditionally localized to the somatosensory cortex and considered to reflect somatosensory activity (Hari, 2006). However, in light of recent findings in perception (Cuellar et al., 2012; Tamura et al., 2012; Pineda et al., 2013), there is mounting evidence that it also may reflect auditory feedback. In speech production, this makes sense considering how both auditory and somatosensory integration regions provide feedback to the PMC during speech production. Furthermore, the feedback from the auditory system and somatosensory system are generally consistent during speech production such that, barring perturbation to either modality, SFC models allow for them to often be considered unitarily (Houde and Nagarajan, 2011).

μ-beta during production

During overt production (SylP and WorP), beta μ-ERD is easily explained as a consequence of motor activity. Sensorimotor beta power has been ubiquitously found to suppress to motor activity from effectors including the fingers (Gaetz et al., 2010), wrist (Alegre et al., 2003), shoulder (Stančák et al., 2000), foot (Pfurtscheller and Lopes Da Silva, 1999), and tongue (Crone et al., 1998). However, if SFC models are applied, beta ERD can be cautiously interpreted as an index of PMC activity in the generation of feedforward control to motor effectors and forward internal models (efference copies) to the feedback loop. This interpretation is supported in a recent review (Engel and Fries, 2010; Kilavik et al., 2013), suggesting the difficulty in determining a clear functional role of sensorimotor beta suppression during movement, but that it may reflect sensory and cognitive aspects (e.g., forward modeling) in addition to pure motor processes. That said, one limitation of the current interpretation is the inability of beta ERD to distinguish between motor activity in feedforward (i.e., muscle movements) and feedback (i.e., internal modeling) mechanisms.

Covert speech production

The Img (covert production) produced significantly weaker alpha/beta ERD than the overt production conditions. This condition was incorporated into the design as previous work examining motor imagery and covert speech production had shown patterns of μ suppression and sensorimotor activity similar to overt productions (Pfurtscheller and Lopes Da Silva, 1999; Neuper et al., 2006). However, there is also evidence that responses in these covert conditions have been weaker than in actual overt productions (Neuper et al., 2006). In a recent study, Holler et al. (2013) investigated μ activity to real and imagined hand movements and showed that only 11 of 18 participants produced differences in μ- alpha/beta power when imagining hand movements. Of these 11, two showed μ enhancement rather than the suppression that was shown in the real movement conditions, suggesting variable responses to covert production tasks. The results in the current Img condition showed similar variability, perhaps contraindicating future use of covert production over a large number of repeated trials. This was the only condition in the experiment that required no overt response (either button-press or speech production) and hence, it was impossible to monitor the extent of covert syllable productions that was asked of participants 80 times in this condition.

Early μ-ERD in overt production

Significantly strong μ-ERD (relative to Img) was found as speech was being produced. Weaker μ-ERD was observed in SylP and WorP conditions prior to production and even before the cue to speak. This time period coincided with the preparation of the speech network, during which similar oscillatory activity has been reported (Gehrig et al., 2012; Herman et al., 2013). μ-ERD during this time period prior to production was weaker than expected, especially in light of the findings in the speech perception tasks, and predictions from SFC models (Houde and Nagarajan, 2011). This reduced neural ERD was most likely due to the influence of EMG on overall EEG variance.

THE UTILITY OF ICA IN SPEECH PERCEPTION AND PRODUCTION

μ components were successfully identified from band-pass filtered concatenated EEG data from perception and production conditions. Though the unexplained RV of the average μ ECD was slightly higher than has been found in other studies (e.g., Bowers et al., 2013), the combination of ECD/sLORETA CSD techniques produced a reliable and valid estimate of μ sources within the standard head model that was applied to all ICA data.

In the perception conditions, time-frequency analyses revealed differential contributions from alpha and beta bands of the μ rhythm that contributed to accurate syllable discrimination. μ-alpha/beta ERD was also revealed in speech production synchronized to muscle activity. This pattern of activity had not been described previously and can be interpreted as being consistent with "normal" sensorimotor control in speech production. Future investigations involving auditory or somatosensory speech perturbations (e.g., Bauer et al., 2006; Reilly and Dougherty, 2013) might be expected to reveal differences in alpha/beta ERD in speech production. Similarly, different relative patterns of μ ERS/ERD might be observed in clinical populations with compromised sensorimotor control such as in stuttering (Max et al., 2003; Loucks and De Nil, 2006; Watkins et al., 2008; Hickok et al., 2011; Cai et al., 2014; Connally et al., 2014).

In addition to the positive findings, there was also evidence of drawbacks to using EEG/ICA in production tasks. It was clear that ICA adequately separated neural from non-neural (e.g., myogenic) activity. Had this not been successfully accomplished, μ-ERD/ERS during production would likely have been overwhelmed by the EMG activity. However, it also appears that overall spectral power in μ components was reduced in the production tasks due to a greater proportion of the overall EEG variance that had to be accounted for by EMG activity. Considering motor requirements, strongest μ-ERD (especially beta) would have been expected in production conditions. However, even when at their strongest, spectral powers during production did not exceed those in perception. In addition, only weak μ-ERD was noted in the time period prior to overt production, which was expected to be stronger as the speech networks prepared to articulate. Together, these findings indicate that overall spectral power in production conditions was attenuated. As such, though interesting general patterns of μ-ERD were revealed in speech production, they should be interpreted with caution with respect to their sensitivity and without making reference to function in conditions without motor requirements.

Another limitation in the current methods was the inability to observe μ activity following speech production. Though production targets (i.e., syllables and words) were produced within the time course of trials, EMG activity (e.g., lip movement) persisted past production, such that the epoch length that did not allow for the measurement of beta rebound (i.e., ERS), which is commonly observed following termination of a movement (Kilavik et al., 2013).

CONCLUSIONS AND FUTURE DIRECTIONS

ICA successfully identified μ components in speech perception and production. Time-frequency analyses using ERSP showed real-time changes in alpha/beta power that provided indicators of PMC/sensorimotor contributions to speech-based dorsal stream activity. Localization of μ clusters and ERSP activity in perception and production are in agreement with Rauschecker's (2011) observation that, based on connections to the inferior parietal lobe and posterior auditory cortex, the PMC provides "optimal state estimation" for speech.

Sensitivity of the findings was somewhat reduced in production conditions, most likely due to concomitant myogenic activity. Further applications in speech production might consider additional filtering techniques in addition to ICA. Exquisite temporal resolution combined with economy and availability warrant further use of ICA particularly to understand speech processing in normal and clinical populations. While measuring the temporal dynamics of the μ-rhythm provide rich information about sensorimotor processing, future ICA studies may also investigate multiple components within the speech processing network in addition to measuring connectivity (i.e., coherence) between components.

REFERENCES

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: two Activation Likelihood Estimation (ALE) meta-analyses. *Brain Lang.* 122, 42–54. doi: 10.1016/j.bandl.2012.04.014

Alegre, M., Gurtubay, I. G., Labarga, A., Iriarte, J., Malanda, A., and Artieda, J. (2003). Alpha and beta oscillatory changes during stimulus-induced movement paradigms: effect of stimulus predictability. *Neuroreport* 14, 381–385. doi: 10.1097/01.wnr.0000059624.96928.c0

Alho, J., Sato, M., Sams, M., Schwartz, J. L., Tiitinen, H., and Jaaskelainen, I. P. (2012). Enhanced early-latency electromagnetic activity in the left premotor cortex is associated with successful phonetic categorization. *Neuroimage* 60, 1937–1946. doi: 10.1016/j.neuroimage.2012.02.011

Arnal, L. H. (2012). Predicting "When" using the motor system's beta-band oscillations. *Front. Hum. Neurosci.* 6:225. doi: 10.3389/fnhum.2012.00225

Arnal, L. H., and Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Bai, O., Mari, Z., Vorbach, S., and Hallett, M. (2005). Asymmetric spatiotemporal patterns of event-related desynchronization preceding voluntary sequential finger movements: a high-resolution EEG study. *Clin. Neurophysiol.* 116, 1213–1221. doi: 10.1016/j.clinph.2005.01.006

Bauer, J. J., Mittal, J., Larson, C. R., and Hain, T. C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: an automatic mechanism for stabilizing voice amplitude. *J. Acoust. Soc. Am.* 119, 2363–2371. doi: 10.1121/1.2173513

Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 25, 60–83. doi: 10.3102/10769986025001060

Bickel, S., Dias, E. C., Epstein, M. L., and Javitt, D. C. (2012). Expectancy-related modulations of neural oscillations in continuous performance tasks. *Neuroimage* 62, 1867–1876. doi: 10.1016/j.neuroimage.2012.06.009

Bidet-Caulet, A., Barbe, P. G., Roux, S., Viswanath, H., Barthelemy, C., Bruneau, N., et al. (2012). Dynamics of anticipatory mechanisms during predictive context processing. *Eur. J. Neurosci.* 36, 2996–3004. doi: 10.1111/j.1460-9568.2012.08223.x

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., and Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nat. Neurosci.* 7, 295–301. doi: 10.1038/nn1198

Blockcolsky, V., Frazer, J., and Frazer, D. (2008). *40,000 Selected Words*. San Antonio, TX: Communication Skill Builders.

Bowers, A., Saltuklaroglu, T., Harkrider, A., and Cuellar, M. (2013). Suppression of the mu rhythm during speech and non-speech discrimination revealed by independent component analysis: implications for sensorimotor integration in speech processing. *PLoS ONE* 8:e72024. doi: 10.1371/journal.pone.0072024

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *J. Acoust. Soc. Am.* 103, 3153–3161. doi: 10.1121/1.423073

Burton, M. W., and Small, S. L. (2006). Functional neuroanatomy of segmenting speech and nonspeech. *Cortex* 42, 644–651. doi: 10.1016/S0010-9452(08)70400-3

Burton, M. W., Small, S. L., and Blumstein, S. E. (2000). The role of segmentation in phonological processing: an fMRI investigation. *J. Cogn. Neurosci.* 12, 679–690. doi: 10.1162/089892900562309

Cai, S., Beal, D. S., Ghosh, S. S., Guenther, F. H., and Perkell, J. S. (2014). Impaired timing adjustments in response to time-varying auditory perturbation during connected speech production in persons who stutter. *Brain Lang.* 129, 24–29. doi: 10.1016/j.bandl.2014.01.002

Callan, A. M., Callan, D. E., Tajima, K., and Akahane-Yamada, R. (2006). Neural processes involved with perception of non-native durational contrasts. *Neuroreport* 17, 1353–1357. doi: 10.1097/01.wnr.0000224774.66904.29

Callan, D., Callan, A., Gamez, M., Sato, M. A., and Kawato, M. (2010). Premotor cortex mediates perceptual performance. *Neuroimage* 51, 844–858. doi: 10.1016/j.neuroimage.2010.02.027

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., and Pesaran, B. (2014). Sensory-motor transformations for speech occur bilaterally. *Nature* 507, 94–98. doi: 10.1038/nature12935

Connally, E. L., Ward, D., Howell, P., and Watkins, K. E. (2014). Disrupted white matter in language and motor tracts in developmental stuttering. *Brain Lang.* 131, 25–35. doi: 10.1016/j.bandl.2013.05.013

Costa, A., Strijkers, K., Martin, C., and Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21442–21446. doi: 10.1073/pnas.0908921106

Crawcour, S., Bowers, A., Harkrider, A., and Saltuklaroglu, T. (2009). Mu wave suppression during the perception of meaningless syllables: EEG evidence of motor recruitment. *Neuropsychologia* 47, 2558–2563. doi: 10.1016/j.neuropsychologia.2009.05.001

Crone, N. E., Miglioretti, D. L., Gordon, B., Sieracki, J. M., Wilson, M. T., Uematsu, S., et al. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization. *Brain* 121(Pt. 12), 2271–2299. doi: 10.1093/brain/121.12.2271

Cuellar, M., Bowers, A., Harkrider, A. W., Wilson, M., and Saltuklaroglu, T. (2012). Mu suppression as an index of sensorimotor contributions to speech processing: evidence from continuous EEG signals. *Int. J. Psychophysiol.* 85, 242–248. doi: 10.1016/j.ijpsycho.2012.04.003

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex* 48, 882–887. doi: 10.1016/j.cortex.2011.05.017

D'Ausilio, A., Pulvermuller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

Dell'acqua, R., Sessa, P., Peressotti, F., Mulatti, C., Navarrete, E., and Grainger, J. (2010). ERP evidence for ultra-fast semantic processing in the picture-word interference paradigm. *Front. Psychol.* 1:177. doi: 10.3389/fpsyg.2010.00177

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., and Makeig, S. (2012). Independent EEG sources are dipolar. *PLoS ONE* 7:e30135. doi: 10.1371/journal.pone.0030135

Doyle, L. M., Yarrow, K., and Brown, P. (2005). Lateralization of event-related beta desynchronization in the EEG during pre-cued reaction time tasks. *Clin. Neurophysiol.* 116, 1879–1888. doi: 10.1016/j.clinph.2005.03.017

Engel, A. K., and Fries, P. (2010). Beta-band oscillations–signaling the status quo? *Curr. Opin. Neurobiol.* 20, 156–165. doi: 10.1016/j.conb.2010.02.015

Erbil, N., and Ungan, P. (2007). Changes in the alpha and beta amplitudes of the central EEG during the onset, continuation, and offset of long-duration repetitive hand movements. *Brain Res.* 1169, 44–56. doi: 10.1016/j.brainres.2007.07.014

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *J. Phon.* 14, 3–28.

Gaetz, W., Macdonald, M., Cheyne, D., and Snead, O. C. (2010). Neuromagnetic imaging of movement-related cortical oscillations in children and adults: age predicts post-movement beta rebound. *Neuroimage* 51, 792–807. doi: 10.1016/j.neuroimage.2010.01.077

Gallese, V., Gernsbacher, M. A., Heyes, C., Hickok, G., and Davis, M. H. (2011). Mirror neuron forum. *Pespect. Psychol. Sci.* 71, 1138–1149. doi: 10.1177/1745691611413392

Ganushchak, L. Y., Christoffels, I. K., and Schiller, N. O. (2011). The use of electroencephalography in language production research: a review. *Front. Psychol.* 2:208. doi: 10.3389/fpsyg.2011.00208

Gehrig, J., Wibral, M., Arnold, C., and Kell, C. A. (2012). Setting up the speech production network: how oscillations contribute to lateralized information routing. *Front. Psychol.* 3:169. doi: 10.3389/fpsyg.2012.00169

Geranmayeh, F., Brownsett, S. L., Leech, R., Beckmann, C. F., Woodhead, Z., and Wise, R. J. (2012). The contribution of the inferior parietal cortex to spoken language production. *Brain Lang.* 121, 47–57. doi: 10.1016/j.bandl.2012.02.005

Ghitza, O., Giraud, A. L., and Poeppel, D. (2012). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Front. Hum. Neurosci.* 6:340. doi: 10.3389/fnhum.2012.00340

Giraud, A. L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063

Gladwin, T. E., Lindsen, J. P., and de Jong, R. (2006). Pre-stimulus EEG effects related to response speed, task switching and upcoming response hand. *Biol. Psychol.* 72, 15–34. doi: 10.1016/j.biopsycho.2005.05.005

Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., and Guenther, F. H. (2011). fMRI investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage* 55, 1324–1338. doi: 10.1016/j.neuroimage.2010.12.065

Grabski, K., Tremblay, P., Gracco, V. L., Girin, L., and Sato, M. (2013). A mediating role of the auditory dorsal pathway in selective adaptation to speech: a state-dependent transcranial magnetic stimulation study. *Brain Res.* 1515, 55–65. doi: 10.1016/j.brainres.2013.03.024

Graimann, B., and Pfurtscheller, G. (2006). Quantification and visualization of event-related changes in oscillatory brain activity in the time—frequency domain. *Prog. Brain Res.* 159, 79–97. doi: 10.1016/S0079-6123(06)59006-5

Gramann, K., Gwin, J. T., Bigdely-Shamlo, N., Ferris, D. P., and Makeig, S. (2010). Visual evoked responses during standing and walking. *Front. Hum. Neurosci.* 4:202. doi: 10.3389/fnhum.2010.00202

Grin-Yatsenko, V. A., Baas, I., Ponomarev, V. A., and Kropotov, J. D. (2010). Independent component approach to the analysis of EEG recordings at early stages of depressive disorders. *Clin. Neurophysiol.* 121, 281–289. doi: 10.1016/j.clinph.2009.11.015

Guenther, F. H., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neurolinguist.* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Gwin, J. T., and Ferris, D. (2011). High-density EEG and independent component analysis mixture models distinguish knee contractions from ankle contractions. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2011, 4195–4198. doi: 10.1109/IEMBS.2011.6091041

Gwin, J. T., Gramann, K., Makeig, S., and Ferris, D. P. (2010). Removal of movement artifact from high-density EEG recorded during walking and running. *J. Neurophysiol.* 103, 3526–3534. doi: 10.1152/jn.00105.2010

Haegens, S., Osipova, D., Oostenveld, R., and Jensen, O. (2010). Somatosensory working memory performance in humans depends on both engagement and

disengagement of regions in a distributed network. *Hum. Brain Mapp.* 31, 26–35. doi: 10.1002/hbm.20842

Hari, R. (2006). Action-perception connection and the cortical mu rhythm. *Prog. Brain Res.* 159, 253–260. doi: 10.1016/S0079-6123(06)59017-X

Herman, A. B., Houde, J. F., Vinogradov, S., and Nagarajan, S. S. (2013). Parsing the phonological loop: activation timing in the dorsal speech stream determines accuracy in speech reproduction. *J. Neurosci.* 33, 5439–5453. doi: 10.1523/JNEUROSCI.1472-12.2013

Hickok, G. (2009). The functional neuroanatomy of language. *Phys. Life Rev.* 6, 121–143. doi: 10.1016/j.plrev.2009.06.001

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* 13, 135–145. doi: 10.1038/nrn3158

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Lang. Cogn. Process.* 29, 2–20. doi: 10.1080/01690965.2013.834370

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hirschfeld, G., Jansma, B., Bolte, J., and Zwitserlood, P. (2008). Interference and facilitation in overt speech production investigated with event-related potentials. *Neuroreport* 19, 1227–1230. doi: 10.1097/WNR.0b013e328309ecd1

Holler, Y., Bergmann, J., Kronbichler, M., Crone, J. S., Schmid, E. V., Thomschewski, A., et al. (2013). Real movement vs. motor imagery in healthy subjects. *Int. J. Psychophysiol.* 87, 35–41. doi: 10.1016/j.ijpsycho.2012.10.015

Houde, J. F., and Nagarajan, S. S. (2011). Speech production as state feedback control. *Front. Hum. Neurosci.* 5:82. doi: 10.3389/fnhum.2011.00082

Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* 10, 371–375.

Jensen, O., Kaiser, J., and Lachaux, J. P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends Neurosci.* 30, 317–324. doi: 10.1016/j.tins.2007.05.001

Kilavik, B. E., Zaepffel, M., Brovelli, A., Mackay, W. A., and Riehle, A. (2013). The ups and downs of beta oscillations in sensorimotor cortex. *Exp. Neurol.* 245, 15–26. doi: 10.1016/j.expneurol.2012.09.014

Klimesch, W. (2012). alpha-band oscillations, attention, and controlled access to stored information. *Trends Cogn. Sci.* 16, 606–617. doi: 10.1016/j.tics.2012.10.007

Klimesch, W., Doppelmayr, M., and Hanslmayr, S. (2006). Upper alpha ERD and absolute power: their meaning for memory performance. *Prog. Brain Res.* 159, 151–165. doi: 10.1016/S0079-6123(06)59010-7

Lau, T. M., Gwin, J. T., and Ferris, D. P. (2014). Walking reduces sensorimotor network connectivity compared to standing. *J. Neuroeng. Rehabil.* 11, 14. doi: 10.1186/1743-0003-11-14

Lee, T. W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput.* 11, 417–441. doi: 10.1162/089976699300016719

Leiberg, S., Lutzenberger, W., and Kaiser, J. (2006). Effects of memory load on cortical oscillatory activity during auditory pattern working memory. *Brain Res.* 1120, 131–140. doi: 10.1016/j.brainres.2006.08.066

Leocani, L., Toro, C., Manganotti, P., Zhuang, P., and Hallett, M. (1997). Event-related coherence and event-related desynchronization/synchronization in the 10 Hz and 20 Hz EEG during self-paced movements. *Electroencephalogr. Clin. Neurophysiol.* 104, 199–206. doi: 10.1016/S0168-5597(96)96051-7

Leocani, L., Toro, C., Zhuang, P., Gerloff, C., and Hallett, M. (2001). Event-related desynchronization in reaction time paradigms: a comparison with event-related potentials and corticospinal excitability. *Clin. Neurophysiol.* 112, 923–930. doi: 10.1016/S1388-2457(01)00530-2

Locasto, P. C., Krebs-Noble, D., Gullapalli, R. P., and Burton, M. W. (2004). An fMRI investigation of speech and tone segmentation. *J. Cogn. Neurosci.* 16, 1612–1624. doi: 10.1162/0898929042568433

Loucks, T. M., and De Nil, L. F. (2006). Anomalous sensorimotor integration in adults who stutter: a tendon vibration study. *Neurosci. Lett.* 402, 195–200. doi: 10.1016/j.neulet.2006.04.002

Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004). Mining event-related brain dynamics. *Trends Cogn. Sci.* 8, 204–210. doi: 10.1016/j.tics.2004.03.008

Max, L., Caruso, A. J., and Gracco, V. L. (2003). Kinematic analyses of speech, orofacial nonspeech, and finger movements in stuttering and nonstuttering adults. *J. Speech Lang. Hear. Res.* 46, 215–232. doi: 10.1044/1092-4388(2003/017)

McMenamin, B. W., Shackman, A. J., Greischar, L. L., and Davidson, R. J. (2011). Electromyogenic artifacts and electroencephalographic inferences revisited. *Neuroimage* 54, 4–9. doi: 10.1016/j.neuroimage.2010.07.057

McMenamin, B. W., Shackman, A. J., Maxwell, J. S., Bachhuber, D. R., Koppenhaver, A. M., Greischar, L. L., et al. (2010). Validation of ICA-based myogenic artifact correction for scalp and source-localized EEG. *Neuroimage* 49, 2416–2432. doi: 10.1016/j.neuroimage.2009.10.010

McMenamin, B. W., Shackman, A. J., Maxwell, J. S., Greischar, L. L., and Davidson, R. J. (2009). Validation of regression-based myogenic correction techniques for scalp and source-localized EEG. *Psychophysiology* 46, 578–592. doi: 10.1111/j.1469-8986.2009.00787.x

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Mottonen, R., and Watkins, K. E. (2012). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology* 26, 1103–1118. doi: 10.1080/02687038.2011.619515

Mottonen, R., Dutton, R., and Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cereb. Cortex* 23, 1190–1197. doi: 10.1093/cercor/bhs110

Moulin-Frier, C., and Arbib, M. A. (2013). Recognizing speech in a novel accent: the motor theory of speech perception reframed. *Biol. Cybern.* 107, 421–447. doi: 10.1007/s00422-013-0557-3

Murakami, T., Ugawa, Y., and Ziemann, U. (2013). Utility of TMS to understand the neurobiology of speech. *Front. Psychol.* 4:446. doi: 10.3389/fpsyg.2013.00446

Muthukumaraswamy, S. D., and Johnson, B. W. (2004). Primary motor cortex activation during action observation revealed by wavelet analysis of the EEG. *Clin. Neurophysiol.* 115, 1760–1766. doi: 10.1016/j.clinph.2004.03.004

Neuper, C., Wortz, M., and Pfurtscheller, G. (2006). ERD/ERS patterns reflecting sensorimotor activation and deactivation. *Prog. Brain Res.* 159, 211–222. doi: 10.1016/s0079-6123(06)59014-4

Nystrom, P. (2008). The infant mirror neuron system studied with high density EEG. *Soc. Neurosci.* 3, 334–347. doi: 10.1080/17470910701563665

Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V. S., and Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Res. Cogn. Brain Res.* 24, 190–198. doi: 10.1016/j.cogbrainres.2005.01.014

Obleser, J., and Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cereb. Cortex* 22, 2466–2477. doi: 10.1093/cercor/bhr325

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Onton, J., Westerfield, M., Townsend, J., and Makeig, S. (2006). Imaging human EEG dynamics using independent component analysis. *Neurosci. Biobehav. Rev.* 30, 808–822. doi: 10.1016/j.neubiorev.2006.06.007

Oostenveld, R., and Oostendorp, T. F. (2002). Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. *Hum. Brain Mapp.* 17, 179–192. doi: 10.1002/hbm.10061

Osnes, B., Hugdahl, K., and Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *Neuroimage* 54, 2437–2445. doi: 10.1016/j.neuroimage.2010.09.078

Pascual-Marqui, R. D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. *Methods Find. Exp. Clin. Pharmacol.* 24(Suppl. D), 5–12.

Perrier, P., Ostry, D. J., and Laboissiere, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *J. Speech Hear. Res.* 39, 365–378.

Perry, A., and Bentin, S. (2010). Does focusing on hand-grasping intentions modulate electroencephalogram mu and alpha suppressions? *Neuroreport* 21, 1050–1054. doi: 10.1097/WNR.0b013e32833fcb71

Pfurtscheller, G., and Lopes Da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* 110, 1842–1857. doi: 10.1016/S1388-2457(99)00141-8

Pfurtscheller, G., Stancak, A. Jr., and Neuper, C. (1996). Post-movement beta synchronization. A correlate of an idling motor area? *Electroencephalogr. Clin. Neurophysiol.* 98, 281–293. doi: 10.1016/0013-4694(95)00258-8

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* 11, 105–110. doi: 10.1016/j.tics.2006.12.002

Pineda, J. A. (2005). The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing." *Brain Res. Brain Res. Rev.* 50, 57–68. doi: 10.1016/j.brainresrev.2005.04.005

Pineda, J. A., Grichanik, M., Williams, V., Trieu, M., Chang, H., and Keysers, C. (2013). EEG sensorimotor correlates of translating sounds into actions. *Front. Neurosci.* 7:203. doi: 10.3389/fnins.2013.00203

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

Rau, C., Plewnia, C., Hummel, F., and Gerloff, C. (2003). Event-related desynchronization and excitability of the ipsilateral motor cortex during simple self-paced finger movements. *Clin. Neurophysiol.* 114, 1819–1826. doi: 10.1016/S1388-2457(03)00174-3

Rauschecker, J. P. (2011). An expanded role for the dorsal pathway in sensorimotor control and integration. *Hear. Res.* 27, 16–25. doi: 10.1016/j.heares.2010.09.001

Rauschecker, J. P. (2012). Ventral and dorsal streams in the evolution of speech and language. *Front. Evol. Neurosci.* 4:7. doi: 10.3389/fnevo.2012.00007

Reilly, K. J., and Dougherty, K. E. (2013). The role of vowel perceptual cues in compensatory responses to perturbations of speech auditory feedback. *J. Acoust. Soc. Am.* 134, 1314–1323. doi: 10.1121/1.4812763

Rogalsky, C., Love, T., Driscoll, D., Anderson, S. W., and Hickok, G. (2011). Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system. *Neurocase* 17, 178–187. doi: 10.1080/13554794.2010.509318

Salmelin, R., and Hari, R. (1994). Spatiotemporal characteristics of sensorimotor neuromagnetic rhythms related to thumb movement. *Neuroscience* 60, 537–550. doi: 10.1016/0306-4522(94)90263-1

Sams, M., Mottonen, R., and Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Brain Res. Cogn. Brain Res.* 23, 429–435. doi: 10.1016/j.cogbrainres.2004.11.006

Sato, M., Tremblay, P., and Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7. doi: 10.1016/j.bandl.2009.03.002

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., and Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Curr. Opin. Neurobiol.* 20, 172–176. doi: 10.1016/j.conb.2010.02.010

Scott, S. K., and Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* 26, 100–107. doi: 10.1016/S0166-2236(02)00037-1

Scott, S. K., McGettigan, C., and Eisner, F. (2009). A little more conversation, a little less action–candidate roles for the motor cortex in speech perception. *Nat. Rev. Neurosci.* 10, 295–302. doi: 10.1038/nrn2603

Shackman, A. J., McMenamin, B. W., Slagter, H. A., Maxwell, J. S., Greischar, L. L., and Davidson, R. J. (2009). Electromyogenic artifacts and electroencephalographic inferences. *Brain Topogr.* 22, 7–12. doi: 10.1007/s10548-009-0079-4

Shahin, A. J., Picton, T. W., and Miller, L. M. (2009). Brain oscillations during semantic evaluation of speech. *Brain Cogn.* 70, 259–266. doi: 10.1016/j.bandc.2009.02.008

Shou, G., Ding, L., and Dasari, D. (2012). Probing neural activations from continuous EEG in a real-world task: time-frequency independent component analysis. *J. Neurosci. Methods* 209, 22–34. doi: 10.1016/j.jneumeth.2012.05.022

Simmonds, A. J., Wise, R. J., Collins, C., Redjep, O., Sharp, D. J., Iverson, P., et al. (2014). Parallel systems in the control of speech. *Hum. Brain Mapp.* 35, 1930–1943. doi: 10.1002/hbm.22303

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *J. Neurosci.* 32, 8443–8453. doi: 10.1523/jneurosci.5069-11.2012

Specht, K. (2013). Mapping a lateralization gradient within the ventral stream for auditory speech perception. *Front. Hum. Neurosci.* 7:629. doi: 10.3389/fnhum.2013.00629

Specht, K. (2014). Neuronal basis of speech comprehension. *Hear. Res.* 307, 121–135. doi: 10.1016/j.heares.2013.09.011

Stančák, A. Jr., and Pfurtscheller, G. (1996). Event-related desynchronisation of central beta-rhythms during brisk and slow self-paced finger movements of dominant and nondominant hand. *Brain Res. Cogn. Brain Res.* 4, 171–183. doi: 10.1016/S0926-6410(96)00031-6

Stančák, A. Jr., Feige, B., Lucking, C. H., and Kristeva-Feige, R. (2000). Oscillatory cortical activity and movement-related potentials in proximal and distal movements. *Clin. Neurophysiol.* 111, 636–650. doi: 10.1016/S1388-2457(99)00310-7

Stevens, K. N., and Halle, M. (1967). *Remarks on Analysis by Synthesis and Distinctive Features*. Cambridge, MA: MIT press.

Strijkers, K., Costa, A., and Thierry, G. (2010). Tracking lexical access in speech production: electrophysiological correlates of word frequency and cognate effects. *Cereb. Cortex* 20, 912–928. doi: 10.1093/cercor/bhp153

Stuart, A., Kalinowski, J., Rastatter, M. P., and Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *J. Acoust. Soc. Am.* 111, 2237–2241. doi: 10.1121/1.1466868

Szenkovits, G., Peelle, J. E., Norris, D., and Davis, M. H. (2012). Individual differences in premotor and motor recruitment during speech perception. *Neuropsychologia* 50, 1380–1392. doi: 10.1016/j.neuropsychologia.2012.02.023

Tamura, T., Gunji, A., Takeichi, H., Shigemasu, H., Inagaki, M., Kaga, M., et al. (2012). Audio-vocal monitoring system revealed by mu-rhythm activity. *Front. Psychol.* 3:225. doi: 10.3389/fpsyg.2012.00225

Tian, X., and Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* 1:166. doi: 10.3389/fpsyg.2010.00166

Tian, X., and Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Front. Hum. Neurosci.* 6:314. doi: 10.3389/fnhum.2012.00314

Tourville, J. A., and Guenther, F. H. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Towle, V. L., Bolanos, J., Suarez, D., Tan, K., Grzeszczuk, R., Levin, D. N., et al. (1993). The spatial location of EEG electrodes: locating the best-fitting sphere relative to cortical anatomy. *Electroencephalogr. Clin. Neurophysiol.* 86, 1–6. doi: 10.1016/0013-4694(93)90061-Y

Tran, Y., Craig, A., Boord, P., and Craig, D. (2004). Using independent component analysis to remove artifact from electroencephalographic measured during stuttered speech. *Med. Biol. Eng. Comput.* 42, 627–633. doi: 10.1007/BF02347544

Venezia, J. H., Saberi, K., Chubb, C., and Hickok, G. (2012). Response bias modulates the speech motor system during syllable discrimination. *Front. Psychol.* 3:157. doi: 10.3389/fpsyg.2012.00157

Ventura, M. I., Nagarajan, S. S., and Houde, J. F. (2009). Speech target modulates speaking induced suppression in auditory cortex. *BMC Neurosci.* 10:58. doi: 10.1186/1471-2202-10-58

Watkins, K. E., Smith, S. M., Davis, S., and Howell, P. (2008). Structural and functional abnormalities of the motor system in developmental stuttering. *Brain* 131, 50–59. doi: 10.1093/brain/awm241

Weisz, N., Hartmann, T., Muller, N., Lorenz, I., and Obleser, J. (2011). Alpha rhythms in audition: cognitive and clinical perspectives. *Front. Psychol.* 2:73. doi: 10.3389/fpsyg.2011.00073

Wilsch, A., Henry, M. J., Herrmann, B., Maess, B., and Obleser, J. (2014). Alpha oscillatory dynamics index temporal expectation benefits in working memory. *Cereb. Cortex*. doi: 10.1093/cercor/bhu004. [Epub ahead of print].

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Zaepffel, M., Trachel, R., Kilavik, B. E., and Brochier, T. (2013). Modulations of EEG beta power during planning and execution of grasping movements. *PLoS ONE* **8**:e60060. doi: 10.1371/journal.pone.0060060

Zion Golumbic, E. M., Poeppel, D., and Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain Lang.* 122, 151–161. doi: 10.1016/j.bandl.2011.12.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 March 2014; accepted: 08 June 2014; published online: 10 July 2014.
Citation: Jenson D, Bowers AL, Harkrider AW, Thornton D, Cuellar M and Saltuklaroglu T (2014) Temporal dynamics of sensorimotor integration in speech perception and production: independent component analysis of EEG data. Front. Psychol. ***5****:656. doi: 10.3389/fpsyg.2014.00656*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 01 July 2014
doi: 10.3389/fpsyg.2014.00673

Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language

Spencer D. Kelly[1,2]*, Yukari Hirata[2,3], Michael Manansala[2,3] and Jessica Huang[2,3]

[1] Neuroscience Program, Department of Psychology, Colgate University, Hamilton, NY, USA
[2] Center for Language and Brain, Colgate University, Hamilton, NY, USA
[3] Department of East Asian Languages and Literatures, Colgate University, Hamilton, NY, USA

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Benjamin Straube, Philipps University, Germany
Pia Knoeferle, Bielefeld University, Germany

***Correspondence:**
Spencer D. Kelly, Neuroscience Program, Department of Psychology, Colgate University, 13 Oak Dr., Hamilton, NY 13346, (315) 228-7350, USA
e-mail: skelly@colgate.edu

Co-speech hand gestures are a type of multimodal input that has received relatively little attention in the context of second language learning. The present study explored the role that observing and producing different types of gestures plays in learning novel speech sounds and word meanings in an L2. Naïve English-speakers were taught two components of Japanese—novel phonemic vowel length contrasts and vocabulary items comprised of those contrasts—in one of four different gesture conditions: Syllable Observe, Syllable Produce, Mora Observe, and Mora Produce. Half of the gestures conveyed intuitive information about syllable structure, and the other half, unintuitive information about Japanese mora structure. Within each Syllable and Mora condition, half of the participants only observed the gestures that accompanied speech during training, and the other half also produced the gestures that they observed along with the speech. The main finding was that participants across all four conditions had similar outcomes in two different types of auditory identification tasks and a vocabulary test. The results suggest that hand gestures may not be well suited for learning novel phonetic distinctions at the syllable level within a word, and thus, gesture-speech integration may break down at the lowest levels of language processing and learning.

Keywords: multimodal, gesture, speech, L2, phoneme, vowel length contrast

INTRODUCTION

The present study explored the question of whether vocabulary and auditory learning in a second language (L2) can be aided by different types of multimodal training, particularly, involving observing or imitating different types of bodily actions. The study is guided by the general view that language processing and learning is fundamentally a *whole body* experience. Indeed, as the current special issue highlights, speech is inherently and systematically embedded within a variety of multimodal behaviors—visual, tactile, and proprioceptive—that are not merely peripheral parts of language, but together with speech, holistically *constitute* language (Clark, 1996; Calvert et al., 2004; McNeill, 2005). There is a rich tradition of research exploring this question in the context of how mouth and lip movements contribute to language processing and learning (sparked by the classic work on the McGurk effect), but more recently, researchers have begun to consider the role that other parts of the body play as well. The present study focuses on one such prominent behavior: hand gesture.

Co-speech gestures are spontaneous hand movements that naturally and ubiquitously accompany speech across different ages, languages, and cultures. Researchers have theorized that gesture and speech stem from the same conceptual starting point in language production, and thus form a fundamentally integrated system of communication (McNeill, 1992; Kendon, 2004). In addition to their role in producing language, co-speech gestures play an active role for language comprehension at various linguistic levels, such as pragmatic, semantic, and syntactic levels (Kelly et al., 2010; Hostetter, 2011; Holle et al., 2012), and this integrated relationship manifests in learning and memory as well (Thompson, 1995; Kelly et al., 1999; Feyereisen, 2006; Straube et al., 2009). Moreover, the role of hand gestures is not limited to one's native language, but they assist in adults' L2 learning as well (Quinn-Allen, 1995; Sueyoshi and Hardison, 2005; Kelly et al., 2009). Kelly et al. (2009), for example, examined the role of iconic gestures in L2 vocabulary learning, and found that English-speakers learned Japanese words better when iconic gestures, such as a drinking gesture, accompanied spoken Japanese words, e.g., *nomu* "drink," compared to when those words were presented alone.

Although most of the research on the integration of gesture and speech focuses on higher levels of analysis (e.g., semantic and pragmatic), there is evidence that the two modalities may also be integrated at lower phonological levels as well (Gentilucci, 2003; Bernardis and Gentilucci, 2006; Krahmer and Swerts, 2007; Hubbard et al., 2008; Biau and Soto-Faraco, 2013). For example, Krahmer and Swerts (2007) showed that, when people produced particular words with beat gestures (which convey rhythmic and prosodic information) in a sentence, they produced those specific words with increased duration and increased pitch height. In addition, when the same spoken words were dubbed into video stimuli with or without those beat gestures, listeners-viewers perceived those words to be more acoustically prominent when

presented with the gestures than without them. Moreover, even within the processing of a single word, gestures affect acoustic features of speech. For example, Gentilucci (2003) showed that viewing different sized gestures made toward different sized objects modulated lip aperture and voice peak amplitude of a speaker producing individual syllables of "BA" and "GA." In this way, gestures can have a significant impact on speech production and comprehension—both in a sentence and word context—even at pre-semantic stages of processing.

Returning to the domain of L2 learning, this opens up an interesting new line of inquiry. Given that other types of visual input (e.g., lip and mouth movements) are well known to help with novel L2 speech perception and learning (Hardison, 2003, 2005; Wang et al., 2008; Hirata and Kelly, 2010), it makes sense to ask what role that hand gestures play in this process as well. In one of the only studies on the topic, Hirata and Kelly (2010) examined the role of co-speech gestures in auditory learning of Japanese vowel length contrasts. Vowel length is phonemic in Japanese, e.g., [kedo] "but" with a short vowel [e] vs. [ke:do] "slight degree" with a long vowel [e:], and L2 learners have difficulty distinguishing these vowel length contrasts (Hirata et al., 2007; Tajima et al., 2008). In Hirata and Kelly (2010), English-speaking participants saw videos of Japanese speakers producing Japanese short and long vowels with and without hand gestures that represented the rhythm of those vowels, i.e., the Syllable gesture in **Figure 1**. A short vertical chopping movement was used for a short vowel, and a long horizontal sweeping movement was used for a long vowel[1]. Contrary to their predictions, participants in the speech-gesture condition did not learn to perceive the short/long vowel contrasts any better than those in the speech alone condition. The authors interpreted this result as hand gestures not playing a role at the segmental phonology level, suggesting a lower limit of speech-gesture integration. However, as the authors also pointed out, it is possible that there might be more effective types of gestures and methods of training. Therefore, the present study explored effects of another type of gesture, i.e., the Mora gesture in **Figure 1**.

The mora is a fundamental unit of timing for Japanese, and a series of moras create temporal beats of roughly equal intervals. The mora is like the syllable but is duration sensitive: long vowels are represented in two moras, or two equal rhythmic beats (Ladefoged, 1975; Port et al., 1987; Vance, 1987; Han, 1994). This mora rhythm is counter-intuitive for English speakers because a long vowel is still one syllable, or one beat. However, the Mora gestures that visually represent two short beats, rather than one extended one, for a long vowel might actually help native-English speakers learn this new mora-based rhythmic system, and ultimately help them hear the short and long vowel distinction better. Indeed, given research showing that "mismatching gestures" complement speech to facilitate learning (Goldin-Meadow, 2005), these counter-intuitive Mora gestures, which hint at a different rhythmic concept, may promote learning by highlighting new and useful strategies.

Another consideration for the design of the present study was that, while Hirata and Kelly (2010) asked learners to observe hand gestures, there is a good reason to believe that observing and imitating, i.e., *producing*, gestures oneself may have a more direct impact on learning than just observing them. There is tradition of research, now labeled "embodied cognition," showing that physically producing actions leads to better learning and memory than just observing them alone (Saltz and Donnenwerth-Nolan, 1981; Cohen, 1989), and recent research has demonstrated that producing gestures is better than just observing them in various instructional settings (Goldin-Meadow et al., 2009; Goldin-Meadow, 2014). In the context of learning an L2, researchers have shown that gesture and speech interact during L2 speech production (for reviews, see Gullberg, 2006; Gullberg et al., 2008) and producing gestures plays a facilitative role in the learning process (Asher's, 1969, "Total Physical Response" technique). A more recent study showed that imitating iconic co-speech gestures helps adults to remember the meaning of words in an invented language more than imitating unrelated hand movements (Macedonia et al., 2011).

With specific regard to Japanese vowel length contrasts, Roberge et al. (1996) taught learners of Japanese to produce hand gestures to differentiate Japanese short and long vowels and observed that these gestures helped learners make significant progress in their short and long vowel production. The explanation of this finding was that by extending the muscles of the arm, the motor system of the arm "resonated" with the vocal motor system, and this made it easier to produce the novels sounds after training. Indeed, research in neuroscience has revealed direct and facilitative connections between producing one's own speech sounds and manual gestures, and this link is also systematically related to the comprehension of the same sounds and gestures produced by others (Bernardis and Gentilucci, 2006; Montgomery et al., 2007; Willems and Hagoort, 2007). These findings suggest that although Hirata and Kelly (2010) showed no unique effect of observing gestures on L2 learners' auditory abilities, *producing* them, in contrast, may be more effective in enabling learners to distinguish the vowel length contrasts.

[1]These gestures were described as "beats" in the Kelly and Hirata paper because of their rhythmic properties. However, these gestures could also be described as "metaphoric" because they cross-modally mapped lengths of visual movements onto lengths of auditory distinctions. Although the gestures used in the present study could also be categorized as both of these types of gesture, we will refer to them as metaphorics.

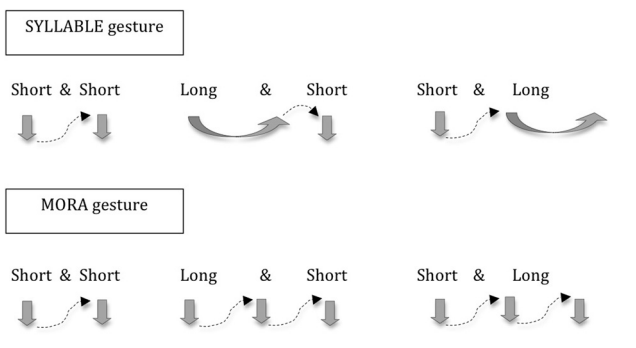


FIGURE 1 | Two types of hand gestures used in the present experiment.

What role do these new methods of training with co-speech hand gestures play in the process of learning to hear difficult L2 sound distinctions and mapping them onto the meaning of new words? Although L2 researchers have traditionally studied phoneme and semantic learning separately, it is important to note that there are close ties between these two abilities when learning an L2. For example, Bundgaard-Nielsen et al. (2011) examined the relationship between the ability to perceive Australian English vowels and vocabulary size by Japanese learners of English, and found that the more accurate their perception, the larger the size of vocabulary. Wong and Perrachione (2007) examined the auditory-vocabulary relationship with learning of Mandarin pseudo words by native English speakers, and found that the learners' ability to attach meaning to the sounds of Mandarin tones (i.e., high-level, rising, and falling) depended on their *initial* auditory ability to identify non-lexical pitch patterns. For Japanese vowel length contrasts, a preliminary finding suggested that a group of learners' very early auditory identification of these contrasts—even before learning meaning of words—enabled greater vocabulary learning as compared with those who learned word meanings first and then were trained to hear these contrasts later (Hirata, 2007).

The ability to hear isolated syllables or words as tested in the above studies, however, might not be a complete measure of learners' ability to perceive fluent sentences. In order to accurately perceive short and long vowels of Japanese, for example, learners must be able to normalize speaking rate of utterances and compare duration of a target vowel with other vowels in a sentence because the duration of "short" or "long" is a relational concept (Hirata, 2004a; Hirata et al., 2007). This *generalized auditory ability* may not necessarily develop if learners are trained only on words in isolation (Hirata, 2004b). Thus, the extent to which various auditory abilities relate to attaching meaning to novel L2 words in a sentence context is still unclear in extant literature.

Given this background, the present study examined effects of multimodal L2 training on auditory abilities through two tasks: the first was *an auditory identification test* in which participants were asked to identify the words they had learned in training (e.g., [seki] "seat" with two short vowels and [se:ki] "century" with one long and one short vowel) vs. untrained words that were similar in syllable compositions but differed in the length of vowels (e.g., [seki:] (nonsense word) with one short and one long vowel). The second task was *an auditory generalization test* in which participants were asked to identify the length of vowels in novel words that they did not hear during training in sentences of different speaking rates. In addition to these two auditory tests, we conducted a vocabulary test consisting of the trained words, e.g., [seki]/[se:ki], and the novel words that differ in vowel length, e.g., [seki:]. This vocabulary test measured how well learners remembered the translation of the trained words, as well as their ability to detect distractor words that differed in length of one of the vowels. The present study examined these three measures in four groups that each went through a different type of multimodal training, (1) Syllable-Observe, in which participants observed the Syllable gesture, (2) Syllable-Produce, in which they observed and produced the Syllable gesture, (3) Mora-Observe, in which they observed the Mora gesture, and (4) Mora-Produce, in which they observed and produced the Mora gesture. The study explored the extent to which these different training types yielded differential results in the above three measures.

In summary, the present study examined the following questions:

(1) Given Goldin-Meadow's (2005) work on gesture-speech "mismatches," does the mora gesture yield a greater auditory and vocabulary learning than the syllable gesture?
(2) Given the literature in embodied learning (Saltz and Donnenwerth-Nolan, 1981; Cohen, 1989), does imitating gestures yield a greater amount of L2 auditory and vocabulary learning than just observing them?
(3) How do effects of the different types of multimodal L2 training manifest in the ability to learn meaning of new words in relation to various auditory abilities, such as differentiating trained and untrained words in a memory task, and identifying short and long vowels in novel words in sentences spoken at different speaking rates?

There are several scenarios as to how our multimodal training may manifest among our different learning measures. For example, if producing gestures contributes to more robust and generalized auditory learning than observing gestures, we predict that participants with the former training would show significantly higher scores on all of these tests. Alternatively, the former group might show significantly higher auditory sentence test scores than the latter group, while both groups score the same on the vocabulary test. There are other possible outcomes as well, but because these were exploratory analyses, we did not have specific a priori predictions about how the four groups' performance would differ on the different dependent measures. However, administering these multiple tests may help tease apart the precise effects of the multimodal input they had received.

METHODS

PARTICIPANTS

Eighty-eight undergraduate students at a liberal arts college in the Northeastern U.S. participated in the study. They were monolingual native speakers of English (males and females) with no knowledge of Japanese language, with an age rage of 18–23. None of these participants had extensive auditory input of Japanese or grew up in bilingual family environments. Participants' formal study of foreign languages included less than 6 years of French, Spanish, German, Italian, Russian, Mandarin Chinese, Arabic, Hebrew, Latin, and Greek. Any participant who had more than 6 years of continuous music training (as screened by a questionnaire) was not included because such musical training is known to affect auditory learning of foreign languages (Sadakata and Sekiyama, 2011). Participants were also screened to be right-handed because the training involved using the right hand in imitating the hand gesture on the computer screen.

Participants were randomly assigned to one of the four training conditions: Syllable-Observe (SO), Syllable-Produce (SP), Mora-Observe (MO), and Mora-Produce (MP) ($n = 22$ in each condition).

OVERALL STRUCTURE OF THE EXPERIMENT

The overall structure of the experiment for all participants was as follows:

- Day 1—an auditory generalization pre-test
- Days 2 and 3—four sessions of training
- Day 4—a vocabulary test and an auditory identification test. (During the auditory identification test, Event Related Potentials (ERPs) were measured, but these results are not reported in the present paper).
- Day 5—an auditory generalization post-test

For Days 1, 4, and 5, all participants took the identical tests. For Days 2 and 3, each participant went through only one of the four types of training. At least 1 day and no more than 3 days separated any 2 Days of the experiment. For example, a participant had to schedule the first day of training (Day 2) at least 1 day but no more than 3 days after the auditory generalization pre-test (Day 1).

TRAINING MATERIALS

Training stimuli were ten pairs of Japanese words that contrasted in length of vowels [e eː o oː u uː] (**Table 1**). The materials included the contrast in the first syllable of the five word pairs and in the second syllable in other five word pairs. To increase variability of stimuli, these words were spoken (in isolation) twice, each with slow and fast speaking rates by two female native speakers of Japanese. The following instructions of the slow and fast speaking rates were given to the speakers: "a slow rate is slower than one's normal rate, clearly enunciating," and "a fast speaking rate is faster than one's normal rate, but still comfortable and accurate." To ensure naturalness, the actual rate of speech was determined by each speaker. A total of 80 audio files (10 word pairs × 2 lengths × 2 repetitions × 2 speakers) were used in the auditory portion of training (Step A in the training procedure section).

In addition, 40 video clips (10 word pairs × 2 lengths × 2 speakers) were created by the same two speakers above to provide the visual dimension of our multimodal training (Steps D, F). For each video clip presenting short vowel words (e.g., [seki] or [joko]), the speaker spoke words and made the hand gesture of two small downward chopping movements. For words with long vowels (e.g., [seːki] or [jokoː]), two types of clips were made, one for the syllable condition and the other for the mora condition. For the syllable condition, the speaker's hand made one horizontal sweep for a long vowel (as in Roberge et al., 1996), followed or preceded by a small downward chopping movement for a short vowel, as they spoke a long-short or short-long word. For the mora condition, in contrast, the speaker's hand made two small downward chopping movements for a long vowel. Thus, they made three short vertical hand movements in long-vowel words (e.g., [seːki] or [jokoː]), and this hand gesture corresponds with the number of moras in those words. Note that gestures for short vowels are identical for the two conditions, with long vowels being the only part where the conditions diverge. Refer to **Figure 1**.

The two native Japanese speakers made 40 total video clips in which the 20 training words (**Table 1**) were spoken and gestured at slow and fast rates. The speaking rate was determined in the same way as when their audio recordings were made. The speakers used their right hand to gesture, and the videos were digitally flipped so that it appeared to be the left hand in order for participants to mirror the gestures they see with their own right hand.

The video clips showed the speaker's face speaking the word and the upper half of the body so that viewers could see hand movements and the face at the same time. In Hirata and Kelly (2010), visual information conveyed through lip movements played a significant role in auditory learning of Japanese vowel length contrasts, and although we did not isolate the lips as a variable in the present study, we wanted to explore the additive role of hand gesture by having both the mouth and hand visible in all conditions.

After auditory and video stimuli were created separately, the audio in the original video was deleted, and new audio clips (spoken without any hand gestures) were dubbed onto the video stimuli. It is known that the acoustic properties of speech are affected by co-speech hand gesture (Krahmer and Swerts, 2007), so it is possible that making the syllable and mora gestures would also alter their speech in subtle ways. By having identical auditory

Table 1 | Training stimuli.

Contrasts in the first syllables			Contrasts in the second syllables		
Word	**Length**	**Meaning**	**Word**	**Length**	**Meaning**
seki	SS	Seat	ɕare	SS	Joke
seːki	LS	Century	ɕareː	SL	Honorarium
kedo	SS	But	goke	SS	Widow
keːdo	LS	Slight degree	gokeː	SL	Word form
toɕo	SS	Book	joko	SS	Side
toːɕo	LS	At the beginning	jokoː	SL	Rehearsal
koʒi	SS	Orphan	iso	SS	Seashore
koːʒi	LS	Construction	isoː	SL	Transport
kuro	SS	Black	ʒiɕu	SS	To turn yourself in
kuːro	LS	Air path	ʒiɕuː	SL	Self-study

Words are written in an International Phonetic Alphabet phonetic transcription. Length SS refers to words with two short vowels; LS, words with a long vowel and a short vowel; SL, words with a short vowel and a long vowel.

information across the four conditions, we assured that any differences in training would be attributed to gesture and not actual differences in the acoustic speech signal.

TRAINING PROCEDURE

Four sessions of training were conducted in Days 2 and 3 (see the overall structure section). At least 1 day and no more than 3 days separated Days 2 and 3 (consistent with the procedure in Hirata and Kelly, 2010). In each training session, participants went through 80 trials, i.e., the 20 words spoken by the two speakers repeated twice in a randomized order. Participants were exposed to only slow rate stimuli in Session 1, only fast rate stimuli in session 2, and both slow and fast rate stimuli in a randomized order in sessions 3 and 4.

The following steps were involved for each of the 80 trials (**Figure 2**):

Step A: Press the space bar to listen to the audio file of the word (produced by one of the two Japanese speakers), and choose one of the three alternatives that corresponded with what they heard: "short-short" (as in [koji] or [joko]), "long-short" (as in [ko:ji]), and "short-long" (as in [joko:]) on the computer screen.

Step B: Watch a video clip in which the speaker said the word along with the accompanying hand gestures (which is indirect feedback regarding the correct answer to participants' response in Step A). Participants in Syllable and Mora conditions saw the syllable and mora gestures, respectively.

Step C: See an English translation of the word written on the screen, e.g., "That means 'construction'."

Step D: See a count down "3," "2," and "1" on the screen for 3 s.

Step E: Watch the same videos as in Step B. The participants in Syllable-Observe and Mora-Observe groups quietly observed the respective videos, and those in Syllable-Produce and Mora-Produce groups mimicked the respective gestures in the videos.

Step F: See the translation of the word again as in Step C.

The participants were instructed to be silent the whole time. Step A (to play the audio and to choose one of the three alternatives) was self-paced, but the other steps were automated by a computer program.

During training, the experimenters monitored the participants through a live video camera to assure that they adhered to their expected tasks in the four conditions. To motivate participants, they were told at the beginning of the first training session that the person who improved most in the test scores would receive a prize.

VOCABULARY TEST (ON DAY 4)

The vocabulary test consisted of 30 words, including the 20 trained words and 10 distractor words. The distractors contained a phonetic composition of consonants and vowels that was identical to the trained words except for the length of the vowels. For example, [seki] "seat" and [se:ki] "century" were trained words, and the distractor was [seki:] (which is a nonsense word because of the length of the two vowels are switched). Materials were made up of (1) the identical audio files used in training, spoken at the slow rate by one of the speakers, and (2) the additional audio files of the distractor words, which were also spoken at the same slow rate by the same speaker. These 30 individual words were organized in a set randomized order, and each word was presented three times with a self-paced pause following each triplet.

The format of the vocabulary test was a free recall task, in which participants were asked to write down the meaning of words in English on a piece of paper, and to write down an "X" if they heard words that sounded similar to the trained words but

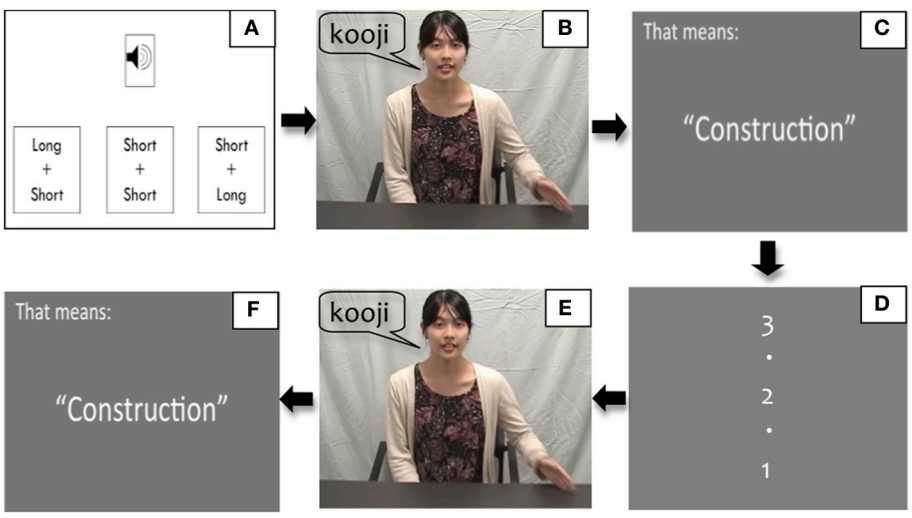


FIGURE 2 | Training steps. (A) Listen to the target word audio, e.g., [ko:ji], and click one of the three alternatives, e.g., "Long + Short"; **(B)** Watch the instructor speaking the target word, e.g., [ko:ji], and showing syllable or mora gestures along with speech; **(C)** See the translation of the target word, e.g., "Construction"; **(D)** See the count down "3, 2, 1"; **(E)** Watch the same video as **(B)** and either observe or produce the respective gesture along with the video; **(F)** See the translation again.

that had different vowel length (i.e., distractor words). The test was self-paced, but participants were told not to go back to previous answers once they moved on to the later trials. The test took about 15 min for each participant to complete.

AUDITORY IDENTIFICATION TEST (ON DAY 4)

Participants' auditory abilities were measured in two tests: an *auditory identification test* (on Day 4) and an *auditory generalization test* (on Days 1 and 5). The purpose of the auditory identification test was to measure participants' ability to immediately recognize the set of words that they had learned in training and to differentiate them from ones that sounded similar but were different from the trained words in terms of length of the vowels. An example of an untrained word would be [seki:] for the word pair [seki] and [se:ki], which was the same as the distractor words in the vocabulary test. The auditory identification test also included untrained words in which both syllables had long vowels, e.g., [se:ki:]. Thus, there was a total of 40 words used in this test, consisting of 20 trained and 20 untrained words.

These 40 words were each presented five times in a randomized order through a speaker. An automated program was created so that the inter-stimulus intervals were at random intervals between 2 and 3 s. The task for participants was a speeded 2-alternative forced identification: participants were asked to press one button as quickly as possible for words they had learned during training, and to press another button for new words that were not trained. This "old-new" format was chosen so that it was compatible with the method of measuring Event Related Potential (ERP) responses at the same time in order to examine how the brain responded to the trained vs. untrained words. Results from the ERP measure, however, will be reported in a separate paper.

AUDITORY GENERALIZATION TESTS (ON DAYS 1 AND 5)

Auditory generalization tests consisted of a pre-test that was conducted before training on Day 1 and a post-test that was conducted after training on Day 5. The purpose of the auditory generalization tests was to measure changes in participants' generalized auditory ability to identify vowel length of *novel* words, rather than to accurately recognize the trained stimuli. Therefore, the words in the generalization tests were all different from those used in training, and each was presented in various carrier sentences produced by a novel female speaker of Japanese who was different from the speakers in the training sessions. The pre- and post-test each contained a total of 120 stimuli. There were 10 target disyllable pairs. For five of the word pairs, the vowel length contrasts were in the first syllable, e.g., [eki] "station" (short + short) vs. [e:ki] "energetic spirit" (long + short), and for the other five word pairs, the contrasts were in the second syllable, e.g., [mizo] "ditch" (short + short) vs. [mizo:] "unprecedented" (short + long). These 20 words were the same for both of the pre-test and the post-test, but were spoken in different carrier sentences, e.g., *sore wa* ___ *da to omou* "I think that is ___." Two carrier sentences used for the pre-test were different from those used for the post-test. Each of these materials was spoken at slow and fast speaking rates.

In order to eliminate any response bias, we needed to match the number of the following three types of words: "short + short," e.g., [eki] and [mizo], "long + short," e.g., [e:ki], and "short + long," e.g., [mizo:]. Of the 120 stimuli in each pre- or post-test, there were 40 "short + short" words (10 words × 2 rates × 2 sentences × 1 repetition), 40 "long + short" words (5 words × 2 rates × 2 sentences × 2 repetitions), and 40 "short + long" words (5 words × 2 rates × 2 sentences × 2 repetitions). By doing this, each item had an equal 33.33% chance of appearing. Half of participants in each of the four conditions heard carrier sentences 1 and 2 at the pre-test, and carrier sentences 3 and 4 at the post-test, and this order was switched for the other half of participants.

For each test, the stimuli described above were randomly presented across word pairs, carrier sentences, and speaking rates. Within each trial, a carrier sentence (e.g., "sore wa ___ da to omou") was written on the computer screen. The participants' task was to listen to varying words inserted in the underlined location and to choose one of three alternatives, i.e., "short + short," "long + short," and "short + long," that matched the vowel length pattern of those varying words. The trials were divided into six blocks for each test, and participants took a short break between blocks. Participants received no feedback on their performance at any time. The task was self-paced, and each test took about 20–30 min to complete. Participants took the auditory post-test within 1–3 days after their final training session.

RESULTS

VOCABULARY SCORES

Although participants learned the vocabulary words in all four conditions (chance performance is 5%), a one-way factorial ANOVA revealed that there were no significant differences across the instruction groups, $F_{(3,\ 84)} = 0.436$, ns. Refer to **Table 2**.

AUDITORY IDENTIFICATION SCORES AND REACTION TIMES

The accuracy rates and reaction times (RTs) were subjected to two separate 2 (trained, untrained) by 4 (SO, SP, MO, MP) mixed ANOVAs[2]. For the accuracy rates, there was no main effect of instruction condition, $F_{(3,\ 79)} = 0.24$, ns, but there was a significant main effect of word type, $F_{(1,\ 79)} = 281.27$, $p < 0.001$, $\eta_p^2 = 0.78$, with trained items ($M = 0.91$, $SD = 0.10$) producing higher accuracy rates than untrained items ($M = 0.59$, $SD = 0.18$) across all instruction conditions. Within each instruction condition, these differences were all significant at the $p < 0.001$ level. There was no significant word type by instruction condition interaction, $F_{(3,\ 79)} = 0.84$, ns. Refer to **Figure 3**.

[2]Five participants were excluded from this analysis because of technical difficulties with the program collecting the error rates and RTs.

Table 2 | Vocabulary scores across the four instruction conditions.

	Syllable observe	Syllable produce	Mora observe	Mora produce
Vocabulary score	0.73 (0.28)	0.67 (0.31)	0.77 (0.23)	0.72 (0.33)

The numbers are proportion correctly recalled, followed by the SDs (in parentheses).

For the RTs, there was a main effect of instruction condition, $F_{(3, 79)} = 3.32$, $p = 0.024$, $\eta_p^2 = 0.11$. Visual inspection of the data suggested that the two Mora conditions ($M = 1633$ ms, $SD = 169$ ms) produced slower RTs than the two Syllable conditions ($M = 1512$ ms, $SD = 204$ ms), and this difference was significant $F_{(1, 81)} = 8.68$, $p = 0.004$, $\eta_p^2 = 0.10$. In addition, there was a significant main effect of word type, $F_{(1, 79)} = 507.06$, $p < 0.001$, $\eta_p^2 = 0.86$, with trained items ($M = 1401$ ms, $SD = 180$ ms) producing faster RTs than untrained items ($M = 1745$ ms, $SD = 233$ ms) across all instruction conditions. Within each instruction condition, these differences were all significant at the $p < 0.001$ level. There was no significant word type by instruction condition interaction, $F_{(3, 79)} = 1.83$, ns. Refer to **Figure 4**.

AUDITORY GENERALIZATION SCORES

The scores on the auditory generalization test were subjected to two a 2 (pre-test, post-test) by 4 (SO, SP, MO, MP) mixed ANOVA. There was a main effect of test time, $F_{(1, 84)} = 66.34$, $p < 0.001$, $\eta_p^2 = 0.44$, with participants in all instruction groups improving from pre- ($M = 0.70$, $SD = 0.15$) to post-test ($M = 0.80$, $SD = 0.14$). Within each instruction condition, these differences were all significant at the $p < 0.001$ level. However, there was no significant main effect of instruction, $F_{(3, 84)} = 0.02$, ns, or interaction of test time and instruction condition, $F_{(3, 84)} = 0.04$, ns. Refer to **Table 3**.

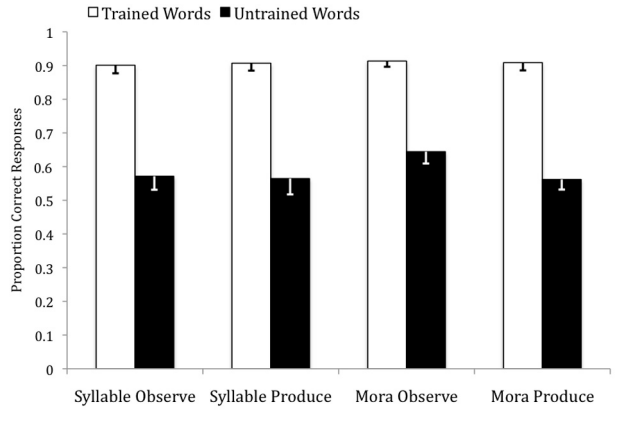


FIGURE 3 | Accuracy rates for trained and untrained words across the four instruction conditions in the auditory identification test.

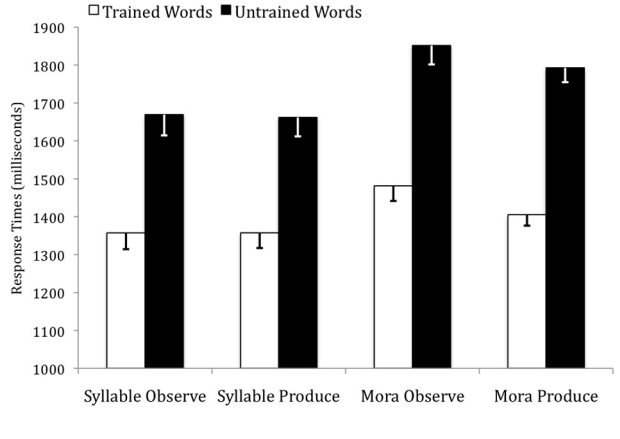


FIGURE 4 | Response times for trained and untrained words across the four instruction conditions in the auditory identification test.

CORRELATIONS AMONG VOCABULARY AND AUDITORY SCORES

In order to investigate whether participants actually applied their auditory learning to performing the vocabulary task, we ran correlations among the vocabulary scores, RTs and accuracy scores for the auditory identification test, and the pre- and post-test auditory generalization scores. In general, vocabulary scores were positively correlated with almost all measures of auditory processing. In particular, note that performance in the auditory identification test accounts for much variance (over 40% for accurately identifying trained items) in the vocabulary performance. Also note that the auditory generalization scores, to a lesser extent, accounts for a sizeable portion of that variance as well[3]. Although both the pre- and post-test both account for significant variance in vocabulary performance (~25 and 20%, respectively), the post-test accounts for significantly more variance when comparing betas in a multiple-regression analysis, pre-test beta: 5.42, $t = 1.01$, ns; post-test beta: 16.39, $t = 2.82$, $p = 0.006$. Refer to **Table 4**.

DISCUSSION

LIMITED ROLE OF HAND GESTURES

We did not find support for our first two predictions in any of the three sets of dependent measures. In none of our measures—vocabulary, auditory identification, and auditory generalization—did the mora and produce conditions outperform the syllable and observe conditions, respectively. This null finding is interesting in light of the fact that our phoneme and vocabulary training was, overall, highly effective. Participants learned vocabulary at a high rate (roughly 70% correct recall), far exceeding chance performance, and all groups improved from pre- to post-test in their ability to distinguish novel phoneme contrasts in the auditory generalization task [similar to previous work, see Hirata et al. (2007)]. Finally, the positive correlations between the two auditory tasks and the vocabulary

[3]When pitted against each other in a multiple regression analysis, accuracy in identifying trained items accounts for significantly more variance than pre- and post-test generalization scores.

Table 3 | Pre- and post-test scores from the generalization auditory test across the four instruction conditions.

Condition	Pre-test	Post-test
Syllable observe	0.70 (0.14)	0.79 (0.13)
Syllable produce	0.70 (0.16)	0.80 (0.13)
Mora observe	0.71 (0.18)	0.80 (0.15)
Mora produce	0.70 (0.14)	0.79 (0.16)

The numbers are proportion correct, followed by the SDs (in parentheses).

Table 4 | Correlation coefficients among the vocabulary scores, response times (RT trained, RT untrained) and accuracy scores for the auditory identification scores (Accuracy trained, Accuracy untrained) and the pre- and post-test auditory generalization scores (Pre-test auditory, Post-test auditory).

	Pre-test auditory	Post-test auditory	Accuracy trained	Accuracy untrained	RT trained	RT untrained
Vocabulary score	0.454**	0.515**	0.640**	0.395**	0.021	0.270*
Pre-test auditory	–	0.764**	0.422**	0.441**	0.101	0.196
Post-test auditory	–	–	0.451**	0.458**	−0.013	0.178
Accuracy trained	–	–	–	0.263*	0.006	0.299**
Accuracy untrained	–	–	–	–	−0.194	−0.218*
RT trained	–	–	–	–	–	0.803**

*$p < 0.05$; **$p < 0.01$ (two-tailed).

task suggest that participants were using their newly acquired phoneme discrimination abilities to remember word meanings, which requires, at a fundamental level, the ability to discriminate long and short vowels. These significant effects also rule out the possibility that we simply did not have enough power to uncover differences across our training conditions. To the contrary, we had moderate-to-large effect sizes in the comparison between pre- and post-tests for the auditory generalization task and very large effect sizes for the RTs and error rates in correctly identifying trained and untrained words in the auditory identification task.

Although one needs to be careful when interpreting null results, the present findings seem to tell a clear story: observing and producing different types of hand gestures *does not* help with learning Japanese long and short vowel distinctions and word meanings comprised of those distinctions. This story is consistent with similar result from a previous study using a comparable training paradigm (Hirata and Kelly, 2010). As described in the introduction, participants in that study were trained to make short and long vowel distinctions in Japanese by observing the same sorts of "syllable gestures" used in this study. The "Observe Syllable" condition in that study (called the Audio + Mouth + Hands condition) produced an improvement of 5% points, which was statistically indistinguishable from a baseline condition of Auditory Only training, which produced a 7% improvement. It is difficult to compare across studies, but it is interesting that the average improvement for all training conditions in the present study (∼9%) was very similar to the Auditory Only improvement in the Hirata and Kelly study.

At first blush, the findings from these two studies are surprising in light of the well-established research—much of it discussed in this special issue—on the benefits of multimodal processing and learning (Calvert et al., 2004). For example, focusing on mouth movements and speech perception, neuroimaging research has shown that visual information from the lips interacts with speech perception at early stages (Klucharev et al., 2003; Besle et al., 2004) and enhances processing in primary visual and auditory cortices compared to visual and auditory input alone (Calvert et al., 1999). With specific regard to L2 learning, research has shown that language learners benefit from instruction that includes speech and visual mouth movements compared to just speech alone (Hardison, 2005; Wang et al., 2008; Hirata and Kelly, 2010).

More recently, researchers have expanded their focus on multimodal communication to include not just the face, but the whole body as well. Indeed, there is growing research on the role of observing and producing hand gestures in language processing and learning (Kelly et al., 2008; Goldin-Meadow, 2014). For example, beat gestures (quick flicks of the hand emphasizing certain words) can change how listeners perceive words (Krahmer and Swerts, 2007; Biau and Soto-Faraco, 2013), and this change in perception is caused by increased activity in auditory brain regions (Hubbard et al., 2008). Moreover, in the context of L2 learning, observing (Kelly et al., 2009) and producing (Macedonia et al., 2011) iconic hand gestures helps to learn and remember new vocabulary in a foreign language.

Given this work on the benefits of multimodal input in language processing and learning, why would observing and producing different types of gesture not help in the present study? We hypothesize that gestures may not be "built for" work at the level in which we applied them. Our gestures were designed to be a visual metaphor of a subtle auditory distinction *within a syllable* at the segmental level. This "within syllable" auditory distinction may be better captured by lip movements, which have a more natural and direct correspondence to the speech they produce. In contrast, gestures may not be easily mapped onto to such small units within a word. Things change when one moves beyond the word level to the sentence level. Indeed, gestures work very well to emphasize the semantically most relevant words within the context of a sentence (Krahmer and Swerts, 2007).

Of course, another explanation for our results is that that gestures do function to make phonemic distinctions within syllables, but just not the phonemic *length* distinctions studied in the present experiment. Recall that the reason English speakers struggle with distinguishing long and short vowels in Japanese is that in English, vowel length is not phonemic—that is, the length of a vowel alone does not change the meaning of a word (Vance, 1987). Indeed, Hirata (2004b) has shown that, for novice English-speaking learners of Japanese, these length distinctions are very hard to learn. Considering this, the auditory contrast may simply be just too foreign and unusual to the novice ear of an English speaker. In contrast, it would be interesting to explore whether gestures play a significant role in learning other types of L2 phonemic contrasts. For example, tones are phonemic in Mandarin, and it would be interesting to examine whether hand gestures that exploit rising and falling space

imitating the tonal contours would help L2 learners to hear the tonal distinction. Another example might be the distinction of different vowel types such as English vowels in *collar* vs. *color*, and it would be interesting to examine whether L2 learners of English would benefit from training with fingers wide open vs. closer together to represent the relative size of the mouth opening.

MORA vs. SYLLABLE GESTURES

One unexpected finding was that for the auditory identification task, people were significantly slower (across the Observe and Produce conditions) to correctly identify trained and untrained words in the two Mora conditions compared to the two Syllable conditions. This finding is notable for a few reasons. First, it demonstrates that there was indeed enough power to uncover significant effects of our training conditions for our different dependent measures. Second, it suggests that, if anything, the mora gestures made the task of identifying trained and untrained words more difficult than syllable gestures—indeed, participants in the two Mora instruction conditions were over 100 ms slower in correctly identifying words than the Syllable condition. This finding provides validation that Mora gestures may indeed be "non-intuitive" to English-speakers, but contrary to this "mismatching" information helping learners, processing them appears to slow learners down. In this way, the Mora gestures act less like the "mismatching" gestures that have been shown to help with learning (Goldin-Meadow, 2014) and more like the "incongruent" gestures that have been shown to slow processing of speech information (Kelly et al., 2010) and disrupt memory for newly learned L2 vocabulary (Kelly et al., 2009).

It would be interesting to explore how more advanced learners of Japanese would react to mora and syllable gestures. Given their more extensive experience with Japanese and better grasp of phonemic distinction between long and short vowels, one might predict that they may have an easier time processing words learned with mora gestures. This raises the interesting possibility that L2 learners may benefit from different types of multimodal input at different stages of learning.

IMPLICATIONS AND CONCLUSION

These findings have important implications for L2 language instruction. We already know from previous research that multimodal input can be very useful when teaching L2 learners novel speech sounds (Hardison, 2003, 2005; Wang et al., 2008; Hirata and Kelly, 2010). These studies have all shown that presenting congruent lip movements with auditory phoneme instruction helps people learn novel phoneme contrasts above and beyond auditory input alone. However, there is evidence that layering *too much* multimodal information onto novel speech sounds may over-load the system and actually produce decrements in perception and learning (Hirata and Kelly, 2010; Kelly and Lee, 2012). For example, Hirata and Kelly (2010) showed that whereas seeing lip movements with speech helped English learners to distinguish Japanese long and short vowels better than speech alone, adding hand gestures to lip and audio training actually removed the positive effects of the mouth.

The present findings add an interesting layer to these studies. When learners have difficulty mapping the meaning of gestures onto novel speech sounds (as with metaphoric gestures conveying information about length of phonemes), it may be wise to eliminate this form of multimodal input from the instruction process, and instead, provide visual input only from the lips and mouth. In contrast, when learners have better mastery with L2 speech sounds, it may be helpful to add gestural input, especially when teaching vocabulary (Quinn-Allen, 1995) and grammar (Holle et al., 2012). So it appears that more multimodal input is not always better in L2 instruction (Hirata and Kelly, 2010; Kelly and Lee, 2012). It will be important to continue this sort of systematic research to carefully demarcate not only what components of second language learning benefit from multimodal input, but also what types of multimodal input optimally enhance those specific components.

Finally, these results are useful in fleshing out claims that gesture and speech constitute an *integrated system* (McNeill, 1992, 2005; Kendon, 2004). For example, McNeill argues that gesture and speech are deeply intertwined and both stem from the same "Growth Point," which he identifies as the conceptual origin of all utterances. When someone gestures, that gesture manifests the most relevant (or "newsworthy," to use McNeill's term) imagistic information contained in that starting point, whereas the speech handles the more traditional functions of language, i.e., the linear, segmentable, and conventional components. Thus, gestures visually highlight information that is conceptually essential to the meaning of an utterance. There is support for this relationship of gesture to speech in the literature on language comprehension (Kelly et al., 2004; Willems et al., 2007; Hostetter, 2011), but the present study suggests that this integrated system may not operate at lower levels of language processing. Perhaps because gestures are so well suited for highlighting semantically relevant information at the utterance level, it is unnatural for them to draw attention to lower level phonemic information *at the segmental timing level*. It will be important for future research on gesture comprehension to more carefully delineate what aspects of gestures form a tightly integrated system with speech—and what aspects do not.

ACKNOWLEDGMENTS

This study was supported by National Science Foundation Grant No. 1052765 given to the first two authors. We thank Carmen Lin, Zach Zhao, April Bailey, and Kristen Weiner at Colgate University for working as a team to prepare stimuli and collect data. We also thank Timothy Collett and Joe Alfonso for programming the training and testing part of our experiment.

REFERENCES

Asher, J. J. (1969). The total physical response approach to second language learning. *Mod. Lang. J.* 53, 3–17. doi: 10.2307/322091

Bernardis, P., and Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia* 44, 178–190. doi: 10.1016/j.neuropsychologia.2005.05.007

Besle, J., Fort, A., Delpuech, C., and Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Biau, E., and Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain Lang.* 124, 143–152. doi: 10.1016/j.bandl.2012.10.008

Bundgaard-Nielsen, R. L., Best, C. T., and Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Stud. Second Lang. Acquis.* 33, 433–461. doi: 10.1017/S0272263111000040

Calvert, G., Spence, C., and Stein, B. E. (eds.). (2004). *The Handbook of Multisensory Processes*. Cambridge, MA: The MIT Press.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., and David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10, 2619–2623. doi: 10.1097/00001756-199908200-00033

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Cohen, R. L. (1989). Memory for action events: the power of enactment. *Educ. Psychol. Rev.* 1, 57–80. doi: 10.1007/BF01326550

Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: their meaning matters. *Eur. J. Cogn. Psychol.* 18, 185–205. doi: 10.1080/09541440540000158

Gentilucci, M. (2003). Grasp observation influences speech production. *Eur. J. Neurosci.* 17, 179–184. doi: 10.1046/j.1460-9568.2003.02438.x

Goldin-Meadow, S. (2005). *Hearing Gesture: How Our Hands Help Us Think*. Cambridge, MA: Harvard University Press.

Goldin-Meadow, S. (2014). How gesture works to change our minds. *Trends Neurosci. Educ.* 3, 4–6. doi: 10.1016/j.tine.2014.01.002

Goldin-Meadow, S., Cook, S. W., and Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychol. Sci.* 20, 267–272. doi: 10.1111/j.1467-9280.2009.02297.x

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition. *Int. Rev. Appl. Linguist. Lang. Teach.* 44, 103–124. doi: 10.1515/IRAL.2006.004

Gullberg, M., de Bot, K., and Volterra, V. (2008). Gestures and some key issues in the study of language development. *Gesture* 8, 149–179. doi: 10.1075/gest.8.2.03gul

Han, M. (1994). Acoustic manifestations of mora timing in Japanese. *J. Acoust. Soc. Am.* 96, 73–82. doi: 10.1121/1.410376

Hardison, D. M. (2003). Acquisition of second-language speech: effects of visual cues, context, and talker variability. *Appl. Psycholinguist.* 24, 495–522. doi: 10.1017/S0142716403000250

Hardison, D. M. (2005). Second-language spoken word identification: effects of perceptual training, visual cues, and phonetic environment. *Appl. Psycholinguist.* 26, 579–596. doi: 10.1017/S0142716405050319

Hirata, Y. (2004a). Effects of speaking rate on the vowel length distinction in Japanese. *J. Phon.* 32, 565–589. doi: 10.1016/j.wocn.2004.02.004

Hirata, Y. (2004b). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *J. Acoust. Soc. Am.* 116, 2384–2394. doi: 10.1121/1.1783351

Hirata, Y. (2007). "A final report on research activities and findings at ATR," in *Advanced Telecommunications Research Institute International* (Kyoto), 1–21.

Hirata, Y., and Kelly, S. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *J. Speech Lang. Hear. Res.* 53, 298–310. doi: 10.1044/1092-4388(2009/08-0243)

Hirata, Y., Whitehurst, E., and Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *J. Acoust. Soc. Am.* 121, 3837–3845. doi: 10.1121/1.2734401

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., and Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Front. Psychol.* 3:74. doi: 10.3389/fpsyg.2012.00074

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychol. Bull.* 137, 297–315. doi: 10.1037/a0022128

Hubbard, A. L., Wilson, S. M., Callan, D. E., and Dapretto, M. (2008). Giving speech a hand: gesture modulates activity in auditory cortex during speech perception. *Hum. Brain Mapp.* 30, 1028–1037. doi: 10.1002/hbm.20565

Kelly, S. D., Barr, D., Church, R. B., and Lynch, K. (1999). Offering a hand to pragmatic understanding: the role of speech and gesture in comprehension and memory. *J. Mem. Lang.* 40, 577–592. doi: 10.1006/jmla.1999.2634

Kelly, S. D., Kravitz, C., and Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain Lang.* 89, 253–260. doi: 10.1016/S0093-934X(03)00335-3

Kelly, S. D., and Lee, A. (2012). When actions speak too much louder than words: gesture disrupts word learning when phonetic demands are high. *Lang. Cogn. Process.* 27, 793–807. doi: 10.1080/01690965.2011.581125

Kelly, S. D., Manning, S., and Rodak, S. (2008). Gesture gives a hand to language and learning: perspectives from cognitive neuroscience, developmental psychology and education. *Lang. Linguist. Compass* 2, 1–20. doi: 10.1111/j.1749-818X.2008.00067.x

Kelly, S. D., McDevitt, T., and Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Lang. Cogn. Process.* 24, 313–334. doi: 10.1080/01690960802365567

Kelly, S. D., Özyürek, A., and Maris, E. (2010). Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychol. Sci.* 21, 260–267. doi: 10.1177/0956797609357327

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge, MA: Cambridge University Press.

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Krahmer, E., and Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *J. Mem. Lang.* 57, 396–414. doi: 10.1016/j.jml.2007.06.005

Ladefoged, P. (1975). *A Course in Phonetics*. Orlando, FL: Harcourt Brace Jovanovich, Inc.

Macedonia, M., Müller, K., and Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Hum. Brain Mapp.* 32, 982–998. doi: 10.1002/hbm.21084

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226514642.001.0001

Montgomery, K. J., Isenberg, N., and Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Soc. Cogn. Affect. Neurosci.* 2, 114–122. doi: 10.1093/scan/nsm004

Port, R. F., Dalby, J., and O'Dell, M. (1987). Evidence for mora timing in Japanese. *J. Acoust. Soc. Am.* 81, 1574–1585. doi: 10.1121/1.394510

Quinn-Allen, L. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *Mod. Lang. J.* 79, 521–529. doi: 10.1111/j.1540-4781.1995.tb05454.x

Roberge, C., Kimura, M., and Kawaguchi, Y. (1996). *Pronunciation Training for Japanese: Theory and Practice of the VT Method* (in Japanese; Nihongo no Hatsuon Shidoo: VT-hoo no Riron to Jissai). Tokyo: Bonjinsha.

Sadakata, M., and Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: a cross-linguistics study. *Acta Psychol.* 138, 1–10. doi: 10.1016/j.actpsy.2011.03.007

Saltz, E., and Donnenwerth-Nolan, S. (1981). Does motoric imagery facilitate memory for sentences? A selective interference test. *J. Verb. Learn. Verb. Behav.* 20, 322–332. doi: 10.1016/S0022-5371(81)90472-2

Straube, B., Green, A., Weis, S., and Chatterjee, A. (2009). Memory effects of speech and gesture binding: cortical and hippocampal activation in relation to subsequent memory performance. *J. Cogn. Neurosci.* 21, 821–836. doi: 10.1162/jocn.2009.21053

Sueyoshi, A., and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Lang. Learn.* 55, 661–699. doi: 10.1111/j.0023-8333.2005.00320.x

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., and Munhall, K. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *J. Acoust. Soc. Am.* 123, 397–413. doi: 10.1121/1.2804942

Thompson, L. A. (1995). Encoding and memory for visible speech and gestures: a comparison between young and older adults. *Psychol. Aging* 10:215. doi: 10.1037/0882-7974.10.2.215

Vance, T. J. (1987). *An Introduction to Japanese Phonology*. Albany, NY: State University of New York Press.

Wang, Y., Behne, D., and Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *J. Acoust. Soc. Am.* 124, 1716–1726. doi: 10.1121/1.2956483

Willems, R. M., and Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: a review. *Brain Lang.* 101, 278–289. doi: 10.1016/j.bandl.2007.03.004

Willems, R. M., Özyürek, A., and Hagoort, P. (2007). When language meets action: the neural integration of gesture and speech. *Cereb. Cortex* 17, 2322–2333. doi: 10.1093/cercor/bhl141

Wong, P. C. M., and Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Appl. Psycholinguist.* 28, 565–585. doi: 10.1017/S0142716407070312

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 March 2014; paper pending published: 16 May 2014; accepted: 10 June 2014; published online: 01 July 2014.

*Citation: Kelly SD, Hirata Y, Manansala M and Huang J (2014) Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. Front. Psychol. **5**:673. doi: 10.3389/fpsyg.2014.00673*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 07 August 2014
doi: 10.3389/fpsyg.2014.00868

Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music

Hweeling Lee[1,2]* and Uta Noppeney[1,3]

[1] Cognitive Neuroimaging Group, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2] Memory Dysfunction in Neurodegenerative Diseases, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[3] Computational Neuroscience and Cognitive Robotics Centre, School of Psychology, University of Birmingham, Birmingham, UK

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Mireille Besson, CNRS, France
Argiro Vatakis, Cognitive Systems Research Institute, Athens, Greece

***Correspondence:**
Hweeling Lee, Memory Dysfunction in Neurodegenerative Diseases, German Center for Neurodegenerative Diseases (DZNE), Ernst-Robert-Curtius Strasse 12, 53117 Bonn, Germany
e-mail: hwee-ling.lee@dzne.de

This psychophysics study used musicians as a model to investigate whether musical expertise shapes the temporal integration window for audiovisual speech, sinewave speech, or music. Musicians and non-musicians judged the audiovisual synchrony of speech, sinewave analogs of speech, and music stimuli at 13 audiovisual stimulus onset asynchronies (±360, ±300 ±240, ±180, ±120, ±60, and 0 ms). Further, we manipulated the duration of the stimuli by presenting sentences/melodies or syllables/tones. Critically, musicians relative to non-musicians exhibited significantly narrower temporal integration windows for both music and sinewave speech. Further, the temporal integration window for music decreased with the amount of music practice, but not with age of acquisition. In other words, the more musicians practiced piano in the past 3 years, the more sensitive they became to the temporal misalignment of visual and auditory signals. Collectively, our findings demonstrate that music practicing fine-tunes the audiovisual temporal integration window to various extents depending on the stimulus class. While the effect of piano practicing was most pronounced for music, it also generalized to other stimulus classes such as sinewave speech and to a marginally significant degree to natural speech.

Keywords: multisensory, temporal synchrony, audiovisual integration, plasticity, speech, music

INTRODUCTION

Music training provides a rich multisensory experience that requires integrating signals from different sensory modalities with motor responses. Thus, the musician's brain provides an ideal model to study experience-dependent plasticity in humans (Munte, 2002; Zatorre et al., 2007). Previous research has shown that musicians develop an enhanced auditory system, both at the structural and functional levels (Schlaug et al., 1995; Munte et al., 2002; Schneider et al., 2005; Hannon and Trainor, 2007; Baumann et al., 2008; Imfeld et al., 2009) that seems to benefit linguistic and non-linguistic skills (Magne et al., 2006; Marques et al., 2007; Moreno et al., 2009; Tzounopoulos and Kraus, 2009; Kraus and Chandrasekaran, 2010). Specifically, musicians proved to be better than non-musicians at segmenting speech from background noise (Parbery-Clark et al., 2013), pitch (Besson et al., 2007), and prosodic tasks (Thompson et al., 2004).

Since practicing a musical instrument for an extensive period of time involves precise timing of several hierarchically organized actions, musical expertise may in particular influence the temporal binding of signals across the senses during perception. Even though sensory signals do not have to be precisely synchronous, they have to co-occur within a certain temporal integration window in order to be integrated into a unified percept (Stein et al., 1993; Spence and Squire, 2003; Noesselt et al., 2007, 2008; Lewis and Noppeney, 2010; Stevenson et al., 2011). Recent studies have shown that the temporal integration window can be narrowed or shifted via long-term musical training (Petrini et al., 2009), short-term perceptual learning (Powers et al., 2009), or short-term audiovisual exposure (Fujisaki et al., 2004). Conversely, it can be widened by exposure to asynchronous stimuli (Navarra et al., 2005).

One critical question is to which extent the impact of musical expertise on audiovisual synchrony perception is specific to the practiced music or whether it generalizes to other stimulus domains. In support of more generic effects, previous studies on auditory processing demonstrated earlier, larger and more robust brainstem responses for musicians relative to non-musicians for both speech and music stimuli (Musacchia et al., 2008; Bidelman and Krishnan, 2010; Bidelman et al., 2011). Moreover, viewing the corresponding videos of the musical instrument in action or facial movements enhanced the temporal and frequency encoding in musicians (Musacchia et al., 2007). Collectively, these results suggest that musical expertise may improve audiovisual processing in a generic fashion at very early processing stages in the brainstem. Based on these results, we may expect that musical expertise fine-tune the temporal integration window generically across multiple stimulus classes such as speech and music.

By contrast, a recent combined psychophysics-fMRI study demonstrated that musicians relative to non-musicians have a significantly narrower temporal integration window for music but not for speech stimuli (Lee and Noppeney, 2011a). Moreover, at the neural level, musicians showed increased audiovisual asynchrony responses and effective connectivity selectively for music but not for speech in a circuitry including the superior temporal

sulcus, the premotor cortex and the cerebellum. These results suggest that music practicing may mold audiovisual temporal binding not only via generic mechanisms of perceptual learning but also via more stimulus-specific mechanisms of sensory-motor learning. More specifically, piano music practicing may fine-tune an internal forward model mapping from action plans specific for piano playing onto visible finger movements and sounds. As this internal forward model furnishes more precise estimates of the relative audiovisual timings of music actions, it sensitizes musicians specifically to audiovisual temporal misalignments of music stimuli. Yet, one may argue that natural speech is not an ideal stimulus class to test whether music expertise transfers from music to other stimulus classes, because both musicians and non-musicians are "speech experts" thereby minimizing any additional effects of musical expertise on audiovisual temporal synchrony perception.

To further investigate whether musical expertise shapes temporal binding of non-music stimuli, we presented 21 musicians and 20 non-musicians participants with natural speech, intelligible sinewave analogs of speech, and piano music stimuli at 13 audiovisual stimulus onset asynchronies (±360, ±300 ±240, ±180, ±120, ±60, and 0 ms) (Dixon and Spitz, 1980; Alais and Burr, 2003; Grant et al., 2004; Zampini et al., 2005; Vatakis and Spence, 2006a,b, 2007, 2008a,b; van Wassenhove et al., 2007; Love et al., 2013). On each trial, participants judged the audiovisual synchrony of natural speech, sinewave speech, and piano music stimuli. We have included these three classes of stimuli to elucidate the main factors that determine whether musical expertise generalizes to other classes of stimuli: Natural speech/sinewave speech and piano music are linked to different motor effectors (mouth vs. hand) and thereby rely on different sensori-motor transformations. By contrast, natural speech and intelligible sinewave speech are identical in the visual facial movements and linguistic representations, but differ in their spectrotemporal structure of the auditory input (Remez et al., 1981; Lee and Noppeney, 2011b; Vroomen and Stekelenburg, 2011; Stekelenburg and Vroomen, 2012; Baart et al., 2014). As sinewave speech is generated by replacing the main speech formants with sinewave analogs, sinewave speech obtains a more musical character. Critically, neither musicians nor non-musicians have been exposed to sinewave speech in their natural environment, so that neither of them are sinewave speech experts. Hence, as with other speech transformations such as rotated speech, both groups should have less precise temporal predictions, and hence, yield a wider temporal integration window for sinewave speech than for piano music or natural speech stimuli (see Maier et al., 2011). These aspects render sinewave speech an ideal stimulus to test for transfer effects from music to other stimulus classes.

Finally, previous studies have demonstrated that humans accumulate statistical information over time for deciding whether auditory and visual signals are synchronous or asynchronous (see Vatakis and Spence, 2006a; Maier et al., 2011). We therefore investigated whether the effect of musical expertise on audiovisual synchrony judgments depends on the stimulus duration by presenting participants with short (piano tones, speech syllables) and long stimuli (piano melodies, speech sentences). In our natural environment human observers are predominantly exposed to connected natural speech and piano music (e.g., melodies), thus, musicians should be familiar with the statistical structure of natural speech and piano music stimuli. Therefore, we expected that the effects of musical training would be more pronounced for long duration stimuli (melodies, speech sentences) as compared to short duration stimuli (piano music tones, speech syllables).

MATERIALS AND METHODS

PARTICIPANTS

Forty-one German native speakers gave informed consent to participate in the study (mean age ± SD = 26 ± 4.9 years). Twenty-one subjects were amateur pianists (mean age ± SD = 24.4 ± 5.1 years) with an average of 16.1 (SD = 5.3) years of experience of piano practicing (mean age of acquisition ± SD = 8.2 ± 2.0 years), and they reported that they practiced the piano for an average of 3.48 (SD = 1.79) hours per week for the last 3 years. In the non-musicians group, all except three subjects (less than 3 months of music training in drums, bass guitar or flute) had no experience with practicing a musical instrument (mean age ± SD = 27.8 ± 4.2 years). The study was approved by the joint human research review committee of the Max Planck Society and the University of Tübingen. A subset of these data (i.e., results for natural speech sentences and melodies) have previously been reported in Lee and Noppeney (2011a, 2014).

DESCRIPTION OF STIMULI

Synchronous audiovisual stimuli were recorded from one speaking actress uttering short sentences or one male hand playing on the piano keyboard (showing one octave) using a camcorder (HVX 200 P, Panasonic Corporation, Osaka, Japan; video at 25 frames per second, PAL 768*567 pixels) for the visual modality and analog recording for the auditory modality (2 channels, 48 kHz). The speech sentences were short neutral statements in German (4–5 words, 7–9 syllables). The music melodies were generated to match the rhythm and number of syllables to those of the speech sentences. The syllables were *"do," "re," "mi," "fa," "so," "la," "di," "to," "bo," "he," "zi," "ka," "lo," "ga," "fi," "po."* The piano music tones were *"do," "re," "mi," "fa," "so," "la," "te," "to."* Supplementary Material shows the list of speech sentences used in the experiment.

The visual and audio recordings were then digitized into MPEG-4 (H.264) format files. The visual file was first cropped to one single complete visual stimulus (speech or music), preceded and followed by 15 frames of neutral facial expression or a still hand image using Adobe Premier Pro (Adobe Systems, San Jose, CA, USA). We added the additional still images to be able to manipulated audiovisual asynchrony without changing the AV length of the stimuli (please see below and Maier et al., 2011).

To transform the auditory modality of natural speech into sinewave speech, the audio tracks were separated from the video tracks. The auditory natural speech was transformed into sinewave speech by replacing the three formants with sinusoid complexes of three sinusoids that were based on the first three vowel formants (www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS). The auditory tracks of sinewave speech were re-combined with the video tracks to create

audiovisual movies of sinewave speech. Four sets of stimuli (24 stimuli per set; 8 stimuli per stimulus class) were created; two sets were stimuli of short duration (i.e., syllables or piano tones) and the other two sets were stimuli of long duration (i.e., sentences or melodies). The sets were counter balanced in time across subjects and across groups.

EXPERIMENTAL DESIGN

The experimental paradigm manipulated: (1) stimulus class: audiovisual speech, sinewave analogs of speech with visual utterance movements of natural speech, audiovisual piano music (i.e., piano music with associated hand movements), (2) stimulus duration: short (single syllables and single piano tones; mean duration $\pm$ $SD = 2.38 \pm 0.37$ s; please note that the duration also include the 15 frame of still images before and after the action sequence), long (sentences and piano melodies; mean duration $\pm$ $SD = 3.56 \pm 0.34$ s), and (3) audiovisual stimulus onset asynchronies (AV-SOA; $\pm$360, $\pm$300, $\pm$240, $\pm$180, $\pm$120, $\pm$60, 0 ms). Positive values indicated that the visual modality was presented first, whereas negative values indicated that the auditory modality was presented first. More specifically, in synchronous stimuli the temporal relationship between the video and the sound track was kept as obtained from recording and thus reflected the natural audiovisual temporal relationship. In other words, it complied with the natural statistics of audiovisual speech or music. Audiovisual asynchronous stimuli were generated by temporally shifting the onset of the auditory track with respect to the video. Moreover, audiovisual synchrony or asynchrony was then determined by the onset of the facial movements and sound rather than the onset of the video (for similar approach and rationale see Maier et al., 2011).

On each trial, subjects judged whether the audiovisual stimuli were synchronous or asynchronous, in an un-speeded fashion. They completed 8 sessions on 2 separate days. Each stimulus was presented 4 times per session in a randomized manner amounting to 2496 (=4 sessions for each stimulus duration $*$ 4 times for each stimulus $*$ 8 stimuli per stimulus class $*$ 3 stimulus classes $*$ 13 AV-SOA) trials. The AV-SOA and stimulus class were randomized in each experiment. The stimuli of short and long duration were presented in separate sessions, and the order was counterbalanced across subjects and days. Prior to the experiment, subjects were presented with all stimuli (2 presentations per stimulus), and then tested on their comprehension of the SWS speech sentences by writing down each sentence that they hear.

EXPERIMENTAL PROCEDURE

The AV-SOA of the separate audio and video files was manipulated using Psychophysics Toolbox version 3 (PTB-3) under Matlab 2007b (MathWorks Inc., MA, USA). Visual stimuli (size 8.89° $*$ 7° visual angle) were projected using a CRT monitor (Sony Trinitron, Tokyo, Japan) at refresh rate of 100 Hz, and subjects' heads were stabilized using a chin rest. Auditory stimuli were presented at ~75 dB SPL via headphones.

DATA ANALYSIS

For each subject and condition, the proportion of synchronous responses (PSR) was computed for each of the 13 AV-SOA levels. To refrain from making any distributional assumptions, the psychometric function was estimated using a non-parametric approach based on local linear fitting methods (Zychaluk and Foster, 2009). The bandwidth for the local quadratic fitting was optimized individually for each subject in a cross-validation procedure. We characterized the psychometric functions by the width of the temporal integration window, as determined by the integral of the psychometric function between -360 and $+360$ ms (after subtracting the difference between one and the maximum from all values of the fitted psychometric function, so that the maximum of all functions was set to one).

To evaluate whether there are any differences in the widths of the temporal integration window between groups, stimulus duration and stimulus class, mixed design ANOVAs were performed with stimulus duration (short, long) and stimulus class (natural speech, sinewave speech, piano music) as within-subject factors, and group (non-musicians, musicians) as a between-subject factor. The results of the ANOVAs are reported after Greenhouse-Geisser correction (when applicable).

RESULTS

After presenting subjects with all stimuli twice (before the main study), we tested them on the comprehension of sinewave speech sentences and syllables. Participants obtained 100% accuracy before the start of the experiment. This ensured that the intelligibility of sinewave speech stimuli could be considered speech-like for the main experiment.

Subjects' PSR for each condition was computed, and psychometric functions were estimated using a non-parametric local quadratic fitting method (Zychaluk and Foster, 2009). **Figure 1** shows the psychometric functions (averaged across subjects) separately for each condition in the musician and non-musician groups. **Figure 2** shows the bar plots of the mean (across subjects' mean) widths of the temporal integration windows for each condition in the musician and non-musician groups. The 2 (group: non-musicians, musicians) $\times$ 2 (stimulus duration: short, long) $\times$ 3 (natural speech, sinewave speech, piano music) mixed design ANOVA on the widths of the temporal integration windows (**Table 1**) revealed a main effect of stimulus duration. Thus, as previously suggested, participants accumulate information over time and thereby obtain more precise temporal estimates for long (i.e., melodies or sentences) relative to short duration stimuli (i.e., piano music tones or syllables) (Maier et al., 2011). Another previous study has reported the opposite finding, i.e., smaller temporal integration windows for syllables as compared to sentences (Vatakis and Spence, 2006a). Vatakis and Spence (2006a) have attributed their results to increased low-level spatiotemporal correlations or increased likelihood of binding attributable to the assumption of "unity" for long relative to short stimuli. However, this previous study differs from the current study in many aspects: (i) they used a temporal order judgment task, (ii) they included only very few stimuli (e.g., only two particular sentences), which makes generalization and interpretation difficult, and (iii) they investigated syllables and sentences in distinct sets of subjects and did not report a formal statistical comparison. For further discussion regarding the issue of stimulus duration, please refer to our previous study that aimed

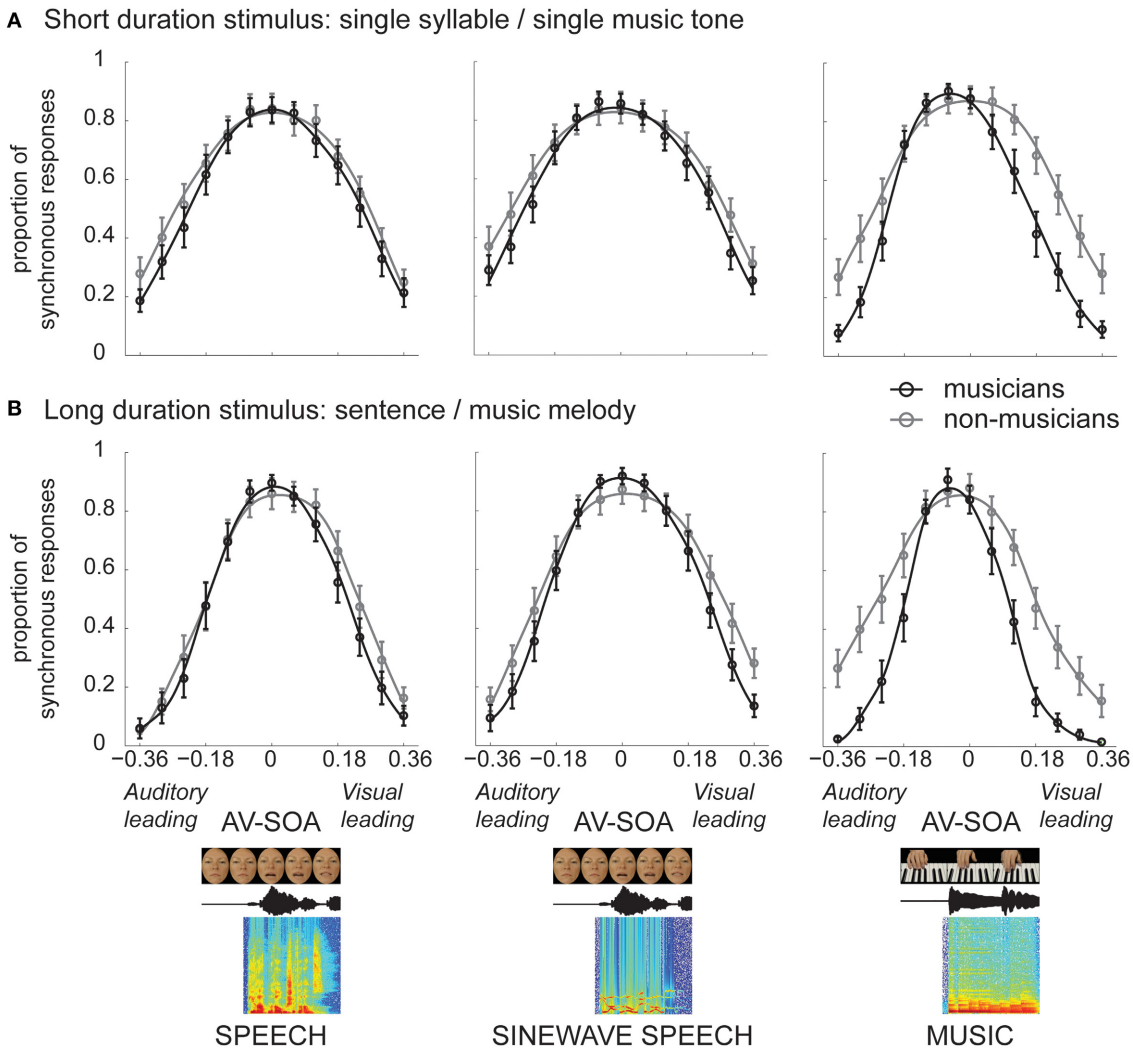


FIGURE 1 | The psychometric functions for speech, sinewave speech, and piano music in non-musicians and musicians for (A) short duration stimulus (syllables or single music tones), and (B) long duration stimulus (sentences or melodies).

to address the influence of stimulus duration on audiovisual temporal integration window (Maier et al., 2011).

Critically, we also observed main effects of stimulus class and group, as well as an interaction between stimulus class and group. As expected, music practice influenced musicians' temporal integration window in a stimulus-dependent fashion and had the strongest effect on piano music stimuli. Thus, as shown in **Figure 3**, the difference in widths of the temporal integration windows for musicians and non-musicians (i.e., the musical expertise effect) was the largest for piano music stimuli. Contrary to our initial hypothesis, we did not observe a significant three-way interaction of stimulus duration, stimulus class and group, a two-way interaction between group and duration, or a two-way interaction between stimulus class and duration. Therefore, we pooled the widths of the temporal integration windows across stimulus duration for natural speech, sinewave speech and piano music, and examined the effect of musical expertise for each stimulus class by computing the difference of the mean widths of the temporal integration windows for musicians relative to non-musicians (i.e., musicians – non-musicians). **Figure 3** depicts the bar plots for the difference (musicians – non-musicians) of the mean widths of temporal integration windows for natural speech, sinewave speech and piano music. Specifically, we tested whether the musical expertise effect (i.e., the difference for musicians – non-musicians) on the widths of temporal integration windows was significantly greater than zero. *Post-hoc* two samples *t*-tests (one-tailed) for each stimulus class revealed that musicians relative to non-musicians exhibited significantly narrower temporal integration windows for sinewave speech [$t_{(39)} = 2.34$, $p = 0.025$; *one-tailed* $p = 0.01$] and piano music [$t_{(39)} = 4.74$, $p < 0.001$], and a marginal significance for natural speech [$t_{(39)} = 1.49$, $p = 0.14$; *one-tailed* $p = 0.07$]. Further, as illustrated in **Figure 3**, we observed a gradient of musical expertise effects for piano music > sinewave speech > natural speech. This

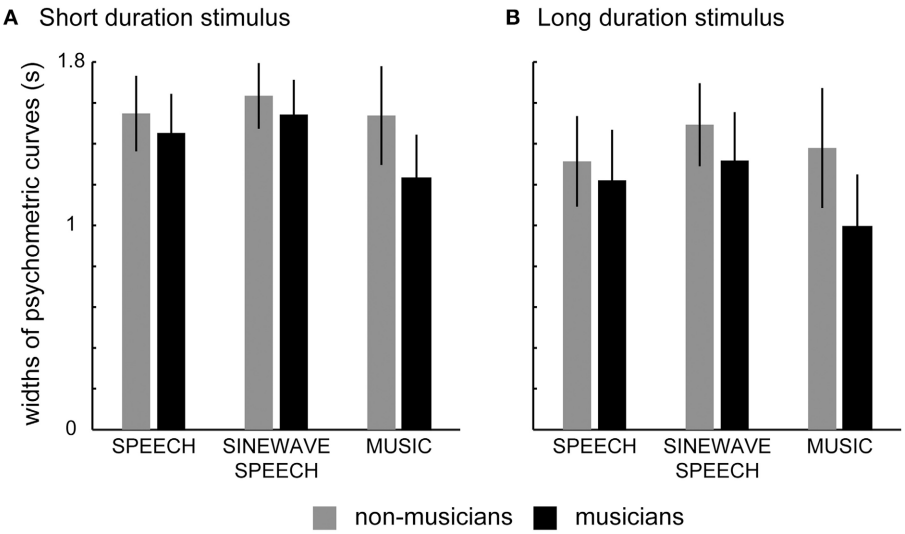


FIGURE 2 | Bar plots showing mean (across subjects' mean) of the widths of the temporal integration windows for speech, sinewave speech, and piano music in non-musicians and musicians for (A) short duration stimulus (syllables or single music tones), and (B) long duration stimulus (sentences or melodies). Error bars represent 1 *SD* (standard deviation).

Table 1 | Results of the mixed ANOVA on the widths of temporal integration windows with stimulus duration (short, long) and stimulus class (speech, sinewave speech, music) as within-subject factors, and group (non-musicians, musicians) as between-subject factor.

Main effects of:		
Group	$\mathbf{F_{(1, 39)} = 10.08}$	$\mathbf{p = 0.003}$
Stimulus duration	$\mathbf{F_{(1, 39)} = 129.5}$	$\mathbf{p < 0.001}$
Stimulus class	$\mathbf{F_{(1.48, 57.8)} = 53.5}$	$\mathbf{p < 0.001}$
Interactions of:		
Group * stimulus duration	$F_{(1, 39)} = 2.20$	$p = 0.146$
Group * stimulus class	$\mathbf{F_{(1.48, 57.8)} = 22.0}$	$\mathbf{p < 0.001}$
Stimulus duration * stimulus class	$F_{(1.46, 56.9)} = 1.60$	$p = 0.215$
Group * stimulus duration * stimulus class	$F_{(1.46, 56.9)} = 1.44$	$p = 0.243$

Significant effects are indicated in bold.

observation was confirmed statistically by *post-hoc* testing for the three interactions that selectively compare the musical expertise effect across two stimulus classes (e.g., musicians – non-musicians for piano music – natural speech). These tests demonstrated that musicians relative to non-musicians exhibited narrower temporal integration windows for piano music > natural speech [$t_{(39)} = 5.41$, $p < 0.001$] and music > sinewave speech [$t_{(39)} = 4.58$, $p < 0.001$], and a marginal significance for sinewave speech > natural speech [$t_{(39)} = 1.51$, $p = 0.14$; *one-tailed* $p = 0.07$]. A one-tailed *t*-test can be adopted, because we would expect a stronger musical expertise effect for sinewave speech than natural speech stimuli (see Introduction).

Collectively, these results demonstrate that the effect of musical expertise was most pronounced for piano music stimuli, and it also generalized to sinewave speech and to a marginally significant extent to natural speech stimuli. However, contrary to our expectations, the musical expertise effect did not depend on stimulus duration. This suggests that even short stimuli provided sufficient statistical structure that enabled musicians to generate more precise estimates of the relative timing of the audiovisual signals.

CORRELATION ANALYSES OF THE WIDTHS OF AUDIOVISUAL TEMPORAL INTEGRATION WINDOWS WITH AGE OF ACQUISITION AND AMOUNT OF PRACTICE

The narrowing of the temporal integration window for musicians may result from innately specified (e.g., genetic) differences between musicians and non-musicians. Alternatively, it may reflect plasticity induced by long-term musical training (Munte, 2002; Zatorre et al., 2007). In the latter case, the narrowing of the temporal integration may depend on the amount of time that musicians spent on piano practicing. Further, the effect of music practice may also interact with neurodevelopment and be most pronounced when children start practicing a musical instrument early in life. In this case, the effect of music practicing should depend on the age at which musicians started piano practicing. Effects of age of acquisition would for instance be observed if piano practicing relies on mechanisms that need to be fine-tuned during sensitive periods in neurodevelopment.

To test whether the narrowing of temporal integration window results from training-induced plasticity, we performed separate correlation analyses testing for a correlation between the width of the psychometric function with (i) age of acquisition or (ii) amount of weekly music practice (in hours) during the past 3 years as predictors. As the widths of temporal integration windows were highly correlated across the different conditions over subjects, we first performed a principal component analysis on the subject-specific widths across all conditions for data reduction. The first component explained 76.3% of the total variance

of all the widths, while the second component explained 9.1% and the third component explained 6.2% of the total variance of all the widths. Thus, as the 2nd component explained only a negligible amount of variance in the data, we extracted and correlated only the first component with age of acquisition and amount of weekly piano music practice during the past 3 years. A significant correlation was found for the first component and amount of weekly piano music practice during the past 3 years [$r_{(21)} = -0.46$, $p = 0.037$] (**Figure 4A**), whereas no significant correlation was found for the first component and age of acquisition [$r_{(21)} = 0.116$, $p = 0.617$] (**Figure 4B**). Specifically, the more the musicians practiced piano, the narrower their temporal integration windows were (i.e., the more sensitive they became to the temporal misalignment of auditory and visual signals).

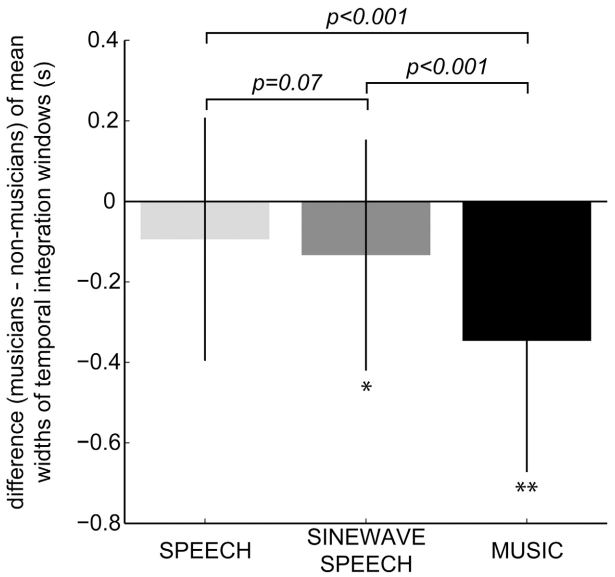


FIGURE 3 | Bar plots showing the musical expertise effect, i.e., difference of the mean widths of the temporal integration windows for musicians vs. non-musicians (musicians – non-musicians; a negative value indicated that musicians relative to non-musicians exhibited a narrower temporal integration window). Error bars represent 1 *SD* (standard deviation). Significance was calculated using one-tailed two samples *t*-tests on the difference of the mean widths of the temporal integration windows for musicians vs. non-musicians (*$p < 0.05$; **$p < 0.001$). Additionally, *p*-values (one-tailed) of one-tailed two samples *t*-tests of the musical expertise effects for sinewave speech > natural speech, piano music > sinewave speech, and piano music > natural speech are shown.

DISCUSSION

Our results demonstrate that long-term music training shapes the temporal integration window in a stimulus-dependent fashion. Musicians, relative to non-musicians, exhibited a narrower temporal integration window predominantly for piano music and to some extent also for sinewave speech with a marginally significant trend for natural speech. Moreover, the amount of weekly piano music practice in the past 3 years correlated with the widths of the temporal integration windows across all stimulus classes. In other words, the more musicians practiced piano in the past 3 years, the more sensitive they became to audiovisual temporal misalignments for natural speech, sinewave speech, and music. Collectively, our results demonstrate that music practice furnishes more precise estimates regarding the relative timings of the audiovisual signals predominantly for music, yet this effect also transferred partly to speech.

Accumulating evidence suggests that music practicing and perceptual learning can influence how human observers temporally bind signals from multiple senses. For instance, a recent psychophysics study demonstrated that musical expertise narrows the temporal integration window for music (Petrini et al., 2009). Yet, this study included only music stimuli. Thus, an unresolved question is to what extent these music or perceptual learning effects are specific to the particular stimulus class trained or whether they can generalize to other stimulus classes. In support

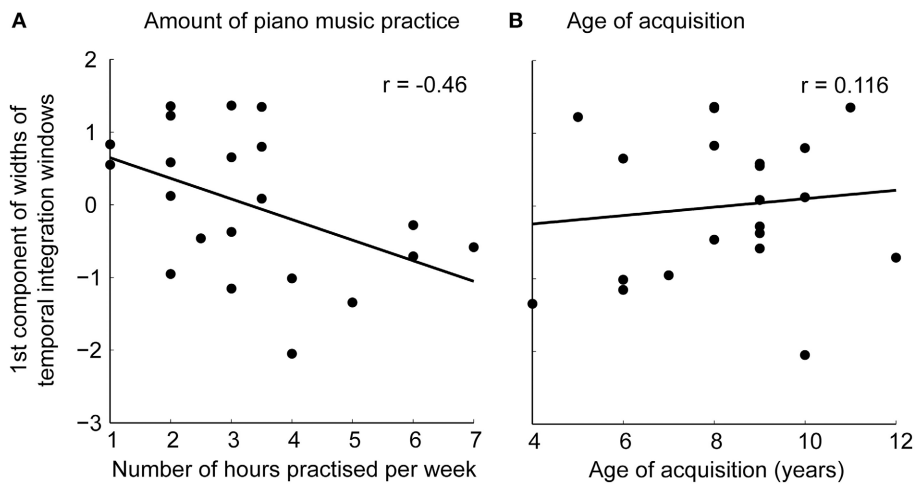


FIGURE 4 | Scatterplots showing correlations of the first component of the widths of temporal integration windows across all stimulus classes with (A) amount of piano music practice over the past 3 years and (B) age of acquisition.

of generic mechanisms of musical expertise, electrophysiological recording demonstrated earlier and larger brain stem responses for musicians relative to non-musicians for both speech and music (Musacchia et al., 2007, 2008; Bidelman and Krishnan, 2010; Bidelman et al., 2011). By contrast, a recent neuroimaging study demonstrated that music practice fine-tunes the temporal integration window predominantly for piano music via engagement of a premotor-cerebellar circuitry (Lee and Noppeney, 2011a).

The current study therefore revisited the question of whether music practice influences audiovisual temporal integration not only of the trained piano music stimuli but also untrained stimulus classes. To this aim, we included natural speech and intelligible sinewave speech signals where the main speech formants have been replaced by sinewave analogs, thereby giving sinewave speech a musical character. Critically, even though the sinewave speech transformation preserved stimulus intelligibility, it introduced a novel mapping between auditory and visual signals. Indeed, as expected, this novel audiovisual mapping made it harder for participants to discriminate between synchronous and asynchronous audiovisual sinewave speech as indicated by a broader integration window for sinewave speech as compared to natural speech (for related findings on rotated speech, see Maier et al., 2011). Thus, the comparison of piano music, sinewave speech and natural speech stimuli enabled us to better characterize to which extent music practice effects transfer to other stimulus classes.

Our results replicate that music expertise shapes temporal binding of audiovisual signals in a stimulus-dependent fashion as indicated by a significant interaction between stimulus class and group. Thus, we observed a gradient of musical expertise effects decreasing from piano music > sinewave speech > natural speech. Nevertheless, the effects of musical expertise on the temporal integration window of other stimulus classes such as sinewave speech or natural speech were still significant. The gradient of musical expertise effect across stimulus classes may be accounted for by two different explanatory frameworks:

First, audiovisual temporal perception may be mediated by only one domain-general mechanism that is engaged by all stimulus classes. Since this domain-general system can be fine-tuned via training to the statistics of a particular stimulus class, the musical expertise effect varies across stimulus classes in a gradual fashion. Thus, pianists would be particularly sensitive to audiovisual asynchronies of piano music stimuli, because the domain-general system has been fine-tuned to the audiovisual temporal statistics of piano music. Yet, transfer effects of musical expertise also emerge, because other stimulus classes can benefit from the fine-tuning of a domain-general system.

Alternatively, the gradient in musical expertise effects may be explained by the concurrent engagement of domain-general and stimulus-specific mechanisms. Domain-general mechanisms have been proposed by a vast number of studies showing musical expertise effects that generalize across music and speech stimuli at the behavioral (Chandrasekaran et al., 2009; Elmer et al., 2012, 2013; Marie et al., 2012; Asaridou and Mcqueen, 2013) or neural level (Musacchia et al., 2008; Bidelman and Krishnan, 2010; Bidelman et al., 2011; Elmer et al., 2012, 2013; Marie et al., 2012). Conversely, we recently showed that music practice sharpens the temporal integration window predominantly for music via premotor-cerebellar circuitry (Lee and Noppeney, 2011a) and proposed that piano practicing may mold the audiovisual temporal integration by training an internal forward model that maps from motor actions (e.g., piano practicing) to its sensory consequences in vision (i.e., finger movements) and audition (e.g., piano sound when hitting the key). Thus, a combination of such a domain-general and a stimulus-dependent sensory-motor mechanism may better explain the transfer of musical expertise effects to other stimulus classes such as sinewave speech in a gradual fashion.

The comparison of musicians and non-musicians cannot resolve ambiguities about whether or not the mechanisms are innately specified or truly reflect experience-dependent plasticity. For instance, amateur musicians may have chosen to practice a musical instrument, because they were inherently better at temporal perception via innate mechanisms. Yet, if musical expertise depends on experience-dependent mechanisms, we would expect that the temporal integration window decrease with the amount of practice. Moreover, if these experience-dependent mechanisms interact with development (e.g., sensitive periods), the integration window should also be influenced by the age at which participants started practicing a musical instrument. Our results demonstrate that indeed the amount of weekly piano practicing in the past 3 years correlates negatively with the musicians' widths of the temporal integration windows—more specifically the first principal component over widths across all conditions. In other words, the more musicians practiced piano, the more sensitive they were to audiovisual temporal misalignments of speech and piano music stimuli. Surprisingly, the age at which musicians started piano practicing did not correlate significantly with the widths of their temporal integration windows. This dissociation suggests that piano practicing shapes audiovisual temporal integration and sensitivity to temporal misalignments via experience-dependent mechanisms that either do not critically interact with neurodevelopment or are bound to sensitive periods in very early development (i.e., before the age of four when the first of our participants started piano practicing). Yet, our results are based on correlative methods. To further substantiate our conclusions, prospective longitudinal studies are required that investigate the change in the temporal integration window as a function of piano music practicing [e.g., 2 (piano practicing vs. other activity) × 2 (before, after training) factorial design].

In conclusion, our results suggest that piano music practicing shapes the temporal integration of audiovisual signals via experience-dependent plasticity. While musical expertise strongly narrows the width of the temporal integration window for piano music, the effect transfers to non-music stimuli such as sinewave speech and a non-significant trend to natural speech. Thus, piano music practicing affects temporal binding either via mechanisms that are specialized predominantly for music but transfer at least in part to other stimulus classes. Alternatively, piano music practicing influences temporal binding of audiovisual signals via multiple mechanisms including stimulus-specific (i.e., specialized for music stimuli) and generic mechanisms (e.g., perceptual learning).

ACKNOWLEDGMENTS

This work is funded by Max Planck Society. We thank Fabian Sinz for creating the music stimuli, Mario Kleiner for help to program the experiment using Matlab with PsychToolbox extensions, Prof. K. Zychaluk and Prof. D. H. Foster for useful discussion regarding the fitting and characterization of the psychometric functions, and Dr. Fabrizio Leo and Dr. Massimiliano Di Luca for comments on the earlier drafts of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00868/abstract

REFERENCES

Alais, D., and Burr, D. (2003). The "flash-lag" effect occurs in audition and cross-modally. *Curr. Biol.* 13, 59–63. doi: 10.1016/S0960-9822(02)01402-1

Asaridou, S. S., and Mcqueen, J. M. (2013). Speech and music shape the listening brain: evidence for shared domain-general mechanisms. *Front. Psychol.* 4:321. doi: 10.3389/fpsyg.2013.00321

Baart, M., Stekelenburg, J. J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 53, 115–121. doi: 10.1016/j.neuropsychologia.2013.11.011

Baumann, S., Meyer, M., and Jancke, L. (2008). Enhancement of auditory-evoked potentials in musicians reflects an influence of expertise but not selective attention. *J. Cogn. Neurosci.* 20, 2238–2249. doi: 10.1162/jocn.2008.20157

Besson, M., Schon, D., Moreno, S., Santos, A., and Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restor. Neurol. Neurosci.* 25, 399–410.

Bidelman, G. M., and Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Res.* 1355, 112–125. doi: 10.1016/j.brainres.2010.07.100

Bidelman, G. M., Krishnan, A., and Gandour, J. T. (2011). Enhanced brainstem encoding predicts musicians' perceptual advantages with pitch. *Eur. J. Neurosci.* 33, 530–538. doi: 10.1111/j.1460-9568.2010.07527.x

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain Lang.* 108, 1–9. doi: 10.1016/j.bandl.2008.02.001

Dixon, N. F., and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception* 9, 719–721.

Elmer, S., Hanggi, J., Meyer, M., and Jancke, L. (2013). Increased cortical surface area of the left planum temporale in musicians facilitates the categorization of phonetic and temporal speech sounds. *Cortex* 49, 2812–2821. doi: 10.1016/j.cortex.2013.03.007

Elmer, S., Meyer, M., and Jancke, L. (2012). Neurofunctional and behavioral correlates of phonetic and temporal categorization in musically trained and untrained subjects. *Cereb. Cortex* 22, 650–658. doi: 10.1093/cercor/bhr142

Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778. doi: 10.1038/nn1268

Grant, K. W., Van Wassenhove, V., and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Commun.* 44, 43–53. doi: 10.1016/j.specom.2004.06.004

Hannon, E. E., and Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends Cogn. Sci.* 11, 466–472. doi: 10.1016/j.tics.2007.08.008

Imfeld, A., Oechslin, M. S., Meyer, M., Loenneker, T., and Jancke, L. (2009). White matter plasticity in the corticospinal tract of musicians: a diffusion tensor imaging study. *Neuroimage* 46, 600–607. doi: 10.1016/j.neuroimage.2009.02.025

Kraus, N., and Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nat. Rev. Neurosci.* 11, 599–605. doi: 10.1038/nrn2882

Lee, H., and Noppeney, U. (2011a). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proc. Natl. Acad. Sci. U.S.A.* 108, E1441–E1450. doi: 10.1073/pnas.1115267108

Lee, H., and Noppeney, U. (2011b). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *J. Neurosci.* 31, 11338–11350. doi: 10.1523/JNEUROSCI.6510-10.2011

Lee, H., and Noppeney, U. (2014). Temporal prediction errors in visual and auditory cortices. *Curr. Biol.* 24, R309–R310. doi: 10.1016/j.cub.2014.02.007

Lewis, R., and Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *J. Neurosci.* 30, 12329–12339. doi: 10.1523/JNEUROSCI.5745-09.2010

Love, S. A., Petrini, K., Cheng, A., and Pollick, F. E. (2013). A psychophysical investigation of differences between synchrony and temporal order judgments. *PLoS ONE* 8:e54798. doi: 10.1371/journal.pone.0054798

Magne, C., Schon, D., and Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children: behavioral and electrophysiological approaches. *J. Cogn. Neurosci.* 18, 199–211. doi: 10.1162/089892906775783660

Maier, J. X., Di Luca, M., and Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 245–256. doi: 10.1037/a0019952

Marie, C., Kujala, T., and Besson, M. (2012). Musical and linguistic expertise influence pre-attentive and attentive processing of non-speech sounds. *Cortex* 48, 447–457. doi: 10.1016/j.cortex.2010.11.006

Marques, C., Moreno, S., Castro, S. L., and Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *J. Cogn. Neurosci.* 19, 1453–1463. doi: 10.1162/jocn.2007.19.9.1453

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., and Besson, M. (2009). Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity. *Cereb. Cortex* 19, 712–723. doi: 10.1093/cercor/bhn120

Munte, T. F. (2002). Brains out of tune. *Nature* 415, 589–590. doi: 10.1038/415589a

Munte, T. F., Altenmuller, E., and Jancke, L. (2002). The musician's brain as a model of neuroplasticity. *Nat. Rev. Neurosci.* 3, 473–478. doi: 10.1038/nrn843

Musacchia, G., Sams, M., Skoe, E., and Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15894–15898. doi: 10.1073/pnas.0701498104

Musacchia, G., Strait, D., and Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hear. Res.* 241, 34–42. doi: 10.1016/j.heares.2008.04.013

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., and Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Brain Res. Cogn. Brain Res.* 25, 499–507. doi: 10.1016/j.cogbrainres.2005.07.009

Noesselt, T., Bonath, B., Boehler, C. N., Schoenfeld, M. A., and Heinze, H. J. (2008). On perceived synchrony-neural dynamics of audiovisual illusions and suppressions. *Brain Res.* 1220, 132–141. doi: 10.1016/j.brainres.2007.09.045

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441. doi: 10.1523/JNEUROSCI.2252-07.2007

Parbery-Clark, A., Strait, D. L., Hittner, E., and Kraus, N. (2013). Musical training enhances neural processing of binaural sounds. *J. Neurosci.* 33, 16741–16747. doi: 10.1523/JNEUROSCI.5700-12.2013

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., et al. (2009). Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Exp. Brain Res.* 198, 339–352. doi: 10.1007/s00221-009-1817-2

Powers, A. C. 3rd., Hillock, A. R., and Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *J. Neurosci.* 29, 12265–12274. doi: 10.1523/JNEUROSCI.3501-09.2009

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–949.

Schlaug, G., Jancke, L., Huang, Y., and Steinmetz, H. (1995). *In vivo* evidence of structural brain asymmetry in musicians. *Science* 267, 699–701.

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., et al. (2005). Structural and functional asymmetry of lateral heschl's gyrus reflects pitch perception preference. *Nat. Neurosci.* 8, 1241–1247. doi: 10.1038/nn1530

Spence, C., and Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Curr. Biol.* 13, R519–R521. doi: 10.1016/S0960-9822(03)00445-7

Stein, B. E., Meredith, M. A., and Wallace, M. T. (1993). The visually responsive neuron and beyond: multisensory integration in cat and monkey. *Prog. Brain Res.* 95, 79–90.

Stekelenburg, J. J., and Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. *Neuropsychologia* 50, 1425–1431. doi: 10.1016/j.neuropsychologia.2012.02.027

Stevenson, R. A., Vanderklok, R. M., Pisoni, D. B., and James, T. W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *Neuroimage* 55, 1339–1345. doi: 10.1016/j.neuroimage.2010.12.063

Thompson, W. F., Schellenberg, E. G., and Husain, G. (2004). Decoding speech prosody: do music lessons help? *Emotion* 4, 46–64. doi: 10.1037/1528-3542.4.1.46

Tzounopoulos, T., and Kraus, N. (2009). Learning to encode timing: mechanisms of plasticity in the auditory brainstem. *Neuron* 62, 463–469. doi: 10.1016/j.neuron.2009.05.002

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., and Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Res.* 1111, 134–142. doi: 10.1016/j.brainres.2006.05.078

Vatakis, A., and Spence, C. (2006b). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neurosci. Lett.* 393, 40–44. doi: 10.1016/j.neulet.2005.09.032

Vatakis, A., and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys.* 69, 744–756. doi: 10.3758/BF03193776

Vatakis, A., and Spence, C. (2008a). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychol.* 127, 12–23. doi: 10.1016/j.actpsy.2006.12.002

Vatakis, A., and Spence, C. (2008b). Investigating the effects of inversion on configural processing with an audiovisual temporal-order judgment task. *Perception* 37, 143–160. doi: 10.1068/p5648

Vroomen, J., and Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: not that special. *Cognition* 118, 75–83. doi: 10.1016/j.cognition.2010.10.002

Zampini, M., Guest, S., Shore, D. I., and Spence, C. (2005). Audio-visual simultaneity judgments. *Percept. Psychophys.* 67, 531–544. doi: 10.3758/BF03193329

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nat. Rev. Neurosci.* 8, 547–558. doi: 10.1038/nrn2152

Zychaluk, K., and Foster, D. H. (2009). Model-free estimation of the psychometric function. *Atten. Percept. Psychophys.* 71, 1414–1425. doi: 10.3758/APP.71.6.1414

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2014; accepted: 21 July 2014; published online: 07 August 2014.
*Citation: Lee H and Noppeney U (2014) Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. Front. Psychol. **5**:868. doi: 10.3389/fpsyg.2014.00868*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 20 May 2014
doi: 10.3389/fpsyg.2014.00422

Atypical audio-visual speech perception and McGurk effects in children with specific language impairment

Jacqueline Leybaert[1]*, Lucie Macchi[2,3], Aurélie Huyse[1], François Champoux[4], Clémence Bayard[1], Cécile Colin[1] and Frédéric Berthommier[5]

[1] Center for Research in Cognition and Neurosciences, Université Libre de Bruxelles, Brussels, Belgium
[2] Ureca, Université de Lille 3, Lille, France
[3] IPSY, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[4] École d'orthophonie et d'audiologie, Université de Montréal, Montréal, QC, Canada
[5] GIPSA-Lab, Université de Grenoble, Grenoble, France

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Nienke Van Atteveldt, VU University Amsterdam, Netherlands
Auli Meronen, Onerva - Centre of Learning and Consulting Services, Finland

***Correspondence:**
Jacqueline Leybaert, Laboratoire Cognition Langage et Développement, Center for Research in Cognition and Neurosciences, Université Libre de Bruxelles, 50 avenue Franklin Roosevelt – CP191, 1050 Brussels, Belgium
e-mail: leybaert@ulb.ac.be

Audiovisual speech perception of children with specific language impairment (SLI) and children with typical language development (TLD) was compared in two experiments using /aCa/ syllables presented in the context of a masking release paradigm. Children had to repeat syllables presented in auditory alone, visual alone (speechreading), audiovisual congruent and incongruent (McGurk) conditions. Stimuli were masked by either stationary (ST) or amplitude modulated (AM) noise. Although children with SLI were less accurate in auditory and audiovisual speech perception, they showed similar auditory masking release effect than children with TLD. Children with SLI also had less correct responses in speechreading than children with TLD, indicating impairment in phonemic processing of visual speech information. In response to McGurk stimuli, children with TLD showed more fusions in AM noise than in ST noise, a consequence of the auditory masking release effect and of the influence of visual information. Children with SLI did not show this effect systematically, suggesting they were less influenced by visual speech. However, when the visual cues were easily identified, the profile of responses to McGurk stimuli was similar in both groups, suggesting that children with SLI do not suffer from an impairment of audiovisual integration. An analysis of percent of information transmitted revealed a deficit in the children with SLI, particularly for the place of articulation feature. Taken together, the data support the hypothesis of an intact peripheral processing of auditory speech information, coupled with a supra modal deficit of phonemic categorization in children with SLI. Clinical implications are discussed.

Keywords: multisensory speech perception, specific language impairment, McGurk effects, audio-visual speech integration, masking release

INTRODUCTION

Children with specific language impairment (SLI) experience difficulties in understanding and producing spoken language, despite normal intelligence, normal hearing, and normal opportunities to learn language. Although linguistic deficits fundamentally characterize SLI (Bishop and Snowling, 2004), theories diverge on the causes of SLI, from grammatical deficit to general or specific limitations in processing capacities (Leonard, 1998, 2004). At the behavioral level, children with SLI are characterized by deficiencies in phonology (Bortolini and Leonard, 2000; Maillart and Parisse, 2006), morphosyntax (Leonard, 1998, 2009) and phonological short-term memory, especially in non-word repetition (Archibald and Gathercole, 2007).

The role of auditory perceptual deficits in explaining the etiology of SLI has been strongly debated. There is much controversy about whether general auditory processing deficits are important in the genesis of specific language disorders (Tallal and Piercy, 1973; Tallal, 1980) or whether the deficit is specific to speech sounds (Mody et al., 1997). Recent work suggest that there are individual differences among children with SLI regarding auditory deficits (Rosen, 2003), and that the deviants may be linked to maturity of auditory processing (Bishop and McArthur, 2004; McArthur and Bishop, 2005). A robust finding in the literature is that even if children with SLI show either no or only subtle speech perception deficits in optimal listening conditions (i.e., in quiet), they exhibit a stronger impairment than children with typical language development (TLD) in speech-in-noise perception. A speech-in-noise deficit in children with SLI has been demonstrated in English (Brady et al., 1983; Robertson et al., 2009; Ferguson et al., 2011) as well as in French (Ziegler et al., 2005, 2009).

Several hypotheses have been advanced to explain the speech-in-noise deficit (Nittrouer et al., 2011). According to a first hypothesis, children with SLI would have an auditory deficit in recovering phonetic structures because of poor sensitivity to formant transitions (Tallal, 1980; Tallal et al., 1993). This idea has been contradicted by several researchers (Sussman, 1993; Bishop et al., 1999; Nittrouer et al., 2011). A second hypothesis is that children with SLI experience more masking of these speech-relevant acoustic properties than children with TLD.

According to Ziegler et al. (2009), children with language problems lack "speech robustness," meaning that they do not have phonological representations as stable as children with TLD. Enhanced masking for speech in children with language problems could be due to those weak representations (Brady et al., 1983; Studdert-Kennedy and Mody, 1995; Johnson et al., 2009; Rosen et al., 2009; Ziegler et al., 2009). The acoustic properties needed for recovering phonetic structure could simply be masked, explaining why phonological representations are so weakly established in the first place (Wright et al., 1997). A third hypothesis is that children with SLI have more difficulties than listeners with TLD at creating well-defined and robust categories in speech as in non-speech. A phonetic category refers to the way various components of the speech signal are combined to form a linguistically meaningful percept. Creation of phonetic categories is related to phonological coding: the language users need to create well-defined categories from sensory information in the signal (Nittrouer et al., 2011).

The difficulties of children with SLI in perceiving speech sounds have been mainly studied in the auditory modality. It appeared that the reception of voicing, place, and manner is impaired in children with SLI compared to age-matched and language-matched children with TLD (Ziegler et al., 2005; but see Collet et al., 2012, for a training of voicing perception in children with SLI).

In face-to-face communication, speech perception is a multimodal process involving both auditory and visual modalities (Sumby and Pollack, 1954; Grant and Seitz, 2000). In noisy contexts, speech detection and comprehension are better in audio-visual conditions (AV), where audition is accompanied by speechreading, than in auditory-only conditions (AO), where only the auditory stimulus is present. During speech perception, auditory and visual cues are merged into a unified percept, a mechanism known as audio-visual (AV) integration. The enhancement afforded by the visual cues in speech-in-noise is largely due to the fact that vision conveys place of articulation, while audition primarily conveys voicing and manner (Summerfield, 1987). The McGurk effect (McGurk and MacDonald, 1976) that occurs when audition and vision provide incongruent tokens illustrates AV integration. For example, when presented with visual velar /ka/ and auditory bilabial /pa/, normally hearing individuals tend to report the illusory fusion alveo-dental /ta/.

Place of articulation is acoustically conveyed by formant transitions, more precisely by the second and third formants, located in high frequencies. The perception of place of articulation is difficult when the acoustic signal is masked by noise (Miller and Nicely, 1955), but is well improved when visual speech cues are added to the signal. When speakers produce /apa/, /ata/, or /aka/, the place of articulation is visually distinguishable by the listener by virtue of the lip movements. Visual information from a talker's face can facilitate speech perception when the environment is less than optimal (Sumby and Pollack, 1954; MacLeod and Summerfield, 1987) or when the listener is hearing impaired (Erber, 1972; Huyse et al., 2012).

Surprisingly, the effect of visual information on speech perception in noise by children with SLI has been little studied up to now. As children with SLI demonstrated a deficit in auditory categorical perception of place of articulation feature (Sussman, 1993; Ziegler et al., 2005; Gerrits and de Bree, 2009), they might take advantage of visual cues, maybe to a greater extent than children with TLD. A few studies examined this question. It appeared that visual articulatory cues influenced adults and children with language impairment to a lesser extend than participants with TLD (Ramirez and Mann, 2005; Norrix et al., 2007; Leybaert and Colin, 2008; Meronen et al., 2013). Ramirez and Mann (2005) compared adults with dyslexia and with auditory neuropathy (AN) to adults with TLD. Participants were presented with natural speech stimuli that were masked with speech-shaped noise at various intensities, in an auditory only (AO), or in an audio-visual (AV) condition. Noise masked the perception of stimuli in AO more in dyslexic and AN participants than in participants with TLD. Patients with AN benefitted from the pairing of visual articulatory cues to auditory stimuli, indicating that their speech perception impairment reflects a peripheral auditory disorder. In contrast, dyslexic participants showed less effective use of visual articulatory cues in identifying masked speech stimuli as well as a lower speechreading capacity relative to control participants. To sum up, language impairment extends beyond the AO modality, and participants with language problems (here: dyslexics) have impoverished AV perception, due to their deficit in speechreading abilities (see also Blau et al., 2010, for a discussion about letter-speech sound integration in developmental dyslexia).

Norrix et al. (2007) presented pre-school children with TLD and with SLI with three syllables /bi/, /di/ and /gi/ in AO, AV congruent and AV incongruent McGurk stimuli (A/bi/ V/gi/ for example). Speechreading ability was not measured. Both groups were at ceiling when asked to identify tokens in AO and AV congruent modalities. A stronger McGurk effect was found for the TLD group compared to the SLI group, indicating that children with SLI were less impacted by the processing of visual speech cues.

Leybaert and Colin (2008) presented French-speaking SLI and TDL children matched for chronological age with video clips of a man speaking /bi/ and /gi/, in optimal listening conditions (no noise). Children with SLI were less likely than TLD children to correctly identify /bi/ and /gi/ syllables in AO as well as in VO modalities. Children with SLI also showed a smaller visual gain (VG), as measured as the improvement of accuracy between AO and AV congruent conditions. When perceiving McGurk incongruent stimuli (e.g., A/gi/V/bi), children with SLI reported more auditory-based responses, fewer visually based responses and fewer combination responses than children with TLD. To sum up, when auditory information is contradicted by visual information such as in McGurk stimuli, children with SLI are less influenced by visual information than children with TLD.

In a recent paper, Meronen et al. (2013) investigated the effect of signal-to-noise ratio (SNR) on the perception of audiovisual speech in 8-year-old children with developmental language disorder and a sample of children with TLD. Performance was measured for /apa/, /ata/, /aka/ presented in AO modality, VO modality, and in AV incongruent (A/p/ V/k/). Three sound intensities (24, 36, and 48 dB) and noise levels (−12, 0, and +6 dB) were used. Both groups achieved similar performance in the AO

condition, but children with developmental language disorders reached lower performances than children with TLD in the VO modality. In response to McGurk stimuli, children with developmental language disorders showed more auditory /p/ responses and less visual /k/ responses than children with TLD. In addition, SNR significantly impacted the proportion of auditory and visual responses in children with TLD, who gave more visual responses when the SNR was more adverse. In contrast, the pattern of responses of children with developmental language disorders was not influenced by SNR. To sum up, the less accurate recognition of visual speech can explain the weaker McGurk effect in the children with developmental language disorders, as well as the lack of impact of SNR on their pattern of auditory and visual responses. This conclusion is in agreement with Norrix et al. (2007) and Leybaert and Colin (2008).

In the current study, we extended the previous investigation by examining the impact of visual cues in the context of a *masking release paradigm*, in school aged children with and without SLI. The release from masking phenomenon refers to the fact that listeners presented with syllables embedded in noise show increased speech intelligibility in fluctuating noise (i.e., modulated in amplitude) compared to stationary noise (Nelson et al., 2003; Füllgrabe et al., 2006). This is an adaptative mechanism since many natural background noises are temporally fluctuating (e.g., surrounding conversations). The masking release phenomenon suggests that listeners are able to "listen in the noise dips" that is, in short temporal minima present in fluctuating noise but absent in stationary noise.

Although children with SLI have lower performances in perceiving auditory syllables masked by either stationary or fluctuating noise, they show an auditory masking release effect comparable to children with TLD (Ziegler et al., 2005). In the present study, we used a new audio-visual masking release paradigm, in which 6 consonants (/apa/, /afa/, /ata/, /asa/, /aʃa/, /aka/) were presented in AO, in VO, and in AV conditions. All stimuli were covered either by stationary or fluctuating noise. AV stimuli were either congruent (e.g., A/apa/ V/apa/) or incongruent (e.g., A/apa/ V/aka/). In previous studies, adults and children with TLD showed larger visual gains when the syllables were masked by stationary noise than when they were masked by fluctuating noise. For incongruent AV stimuli, they gave a majority of visually-based responses when syllables were masked by stationary noise, and more fusions and auditory-based responses when syllables were presented with fluctuating noise (Huyse et al., 2012; Huyse et al., in revision).

As in our previous research, we expected to observe a strengthening of the McGurk effect with fluctuating noise compared to stationary noise in children with TLD. Our main interest was to test whether children with SLI would also show a strengthening of the McGurk effect with fluctuating noise, meaning that their performance would approach that of the TLD children in the conditions of an auditory masking release. We used two types of McGurk stimuli: the plosives A/apa/ V/aka/, giving rise to the fusion /ata/, and the fricatives A/afa/ V/aʃa/, leading to the fusion /asa/ (Berthommier, 2001). The interest of A/afa/ V/aʃa/ is that a dominance of the video responses / ʃ / is observed (Berthommier, 2001; Huyse et al., 2012). If children with SLI recognize /aʃa/ in VO condition, their responses to A/afa/V/aʃa/ would show clear influence of visual information, as it is the case in the TLD children.

Unisensory auditory (AO stimuli) and lipreading (VO stimuli) performances, as well as audio-visual speech perception (AV congruent stimuli) were also measured. In AO, we expected a larger speech-in-noise deficit in children with SLI compared to TLD children, but a similar masking release effect in both groups (Ziegler et al., 2005). In VO, children with SLI would experience more difficulties than TLD children. The visual gain measures the improvement of speech identification in AV compared to AO, due to efficient use of visual cues to recover place of articulation and manner features. Compared to children with TLD, children with SLI would experience less influence of visual cues, and a reduced visual gain.

These hypotheses were tested in two experiments. In Experiment 1, six voiceless consonants were presented in a /aCa/ context, masked by either stationary or amplitude modulated noise (8 Hz and 128 Hz). The stimuli were presented in Audio-only (AO), Visual Only (VO), and Audio-visual (AV) congruent and AV incongruent conditions.

In Experiment 2, larger groups of children with SLI and children with TLD were recruited. Twelve consonants (six voiceless and six voiced) masked either by stationary or by amplitude modulated noise (at 8 Hz) noise were presented in AO, VO, AV congruent conditions, and four McGurk stimuli (two with plosives, two with fricatives) were used. The first aim of Experiment 2 was to replicate the results of Experiment 1 with a large set of consonants. The second aim was to evaluate the specific reception of voicing, place and manner by information transmission (IT) analyses performed on the basis of confusion matrices (Miller and Nicely, 1955). Specifically, we expected an increase of IT in AV compared to AO for the reception of manner and place of articulation, but not for voicing, which has no visible correlate. For the same reason, the percent of IT would be higher than 50% for manner and place of articulation in VO, but around 50% for voicing. Compared to children with TLD, we expected to observe a lower percent of IT across the three features in children with SLI, with a possible enhanced deficit for place of articulation.

EXPERIMENT 1

MATERIAL AND METHOD

Participants

Fifteen French-speaking children with SLI (8 boys) were recruited in special language classes and through an association of parents of children with SLI. The participants met the following criteria: (1) presence of a long-lasting and severe impairment of expressive and/or receptive language, diagnosed as SLI by a neuro-pediatrician in a multi-diciplinary team; (2) no history of hearing loss and no malformation of speech organs; (3) a score > 132 points on the pragmatic component (scales C to G) of the Children's Communication Checklist (Bishop, 1998); (4) a non-verbal IQ > 85 on the French version of the Wechsler Intelligence Scale for children (Wechsler, 1996); and (e) at least 1.5 SD below the age-appropriate mean on the three language tests described below. One child was excluded from our sample, due to the absence of a recent assessment of persistant language

impairment. The final sample included 14 children (7 boys) ranging in age from 8 years 7 months to 14 years 5 months (mean age: 138 months; $SD = 25$ months). All children had measured reading and spelling levels corresponding at least to the end of first grade.

Language assessment tests included: (a) reading aloud of pseudowords and phonically regular and irregular frequent words of the Odedys Test (Jacquier-Roux et al., 2005); (b) Repetition of Difficult Words from the L2MA (Chevrie-Muller et al., 1997); (c) receptive lexical knowledge (EVIP, French version of the PPVT, (Dunn et al., 1993): children have to listen to a word said by the experimenter and to designate the picture corresponding to that word, among four pictures.

A control group of French-speaking children with TLD was recruited. None of them had any history of language or hearing disorders or used hearing aids. Each child with TLD was matched with a child with SLI, based on chronological age and gender. The control group included 14 children (7 boys) ranging in age from 9 years 1 months to 14 years 6 months (mean age: 141 months; SD 25 months). The scores of the children with TLD were within normal limits for the three language tests.

The characteristics of the participants and a summary of the language test scores of the children with SLI and those with TLD are found in **Table 1**. All participants had normal or corrected-to-normal vision and none of them reported any difficulties with viewing the visual stimuli presented in this study.

The project has been reviewed and approved by the University research ethic board. Informed consent was obtained from the parents of all participants, and children provided a verbal acceptance prior to their participation. They were informed that they could interrupt their participation if they felt any problem during the experiment.

Table 1 | Characteristics of children with SLI and of TLD controls—Experiment 1.

	SLI	TLD	Group effect $F_{(1, 26)} =$ p-value
Age in years, months (range)	11.6 (8.7–14.5)	11.9 (9.1–14.6) (9.1–14.6)	*Ns*
Word repetition (*SD*)	20.07 (4.97)	29.86 (0.36)	$F = 54.02$ $p < 0.001$
Vocabulary EVIP (*SD*)	91.14 (22.65)	132.71 (17.28)	$F = 29.81$ $p < 0.001$
Irregular words (*SD*)	7.71 (6.68)	19.07 (1.82)	$F = 37.63$ $p < 0.001$
Regular words (*SD*)	10.29 (6.67)	19.86 (0.53)	$F = 28.62$ $p < 0.001$
Pseudo words (*SD*)	7.07 (5.89)	17.50 (1.74)	$F = 40.34$ $p < 0.001$

Word repetition values indicate number of correct responses (out of 30) on the repetition test taken from the L2MA language battery; Values for Vocabulary indicate raw score on the EVIP test; Irregular Words, Regular Words and Pseudo Words values indicate number of correct responses (out of 20) on the reading tests for frequent items taken from Odedys battery. Standard deviations are in brackets.

Stimuli

Stimuli were composed of vowel-consonant-vowel (VCV) syllables with the consonants /p, t, k, s, f, ʃ/ interposed between two /a/ vowels. A male speaker of French was videotaped while saying these syllables. He was filmed from the bottom of the nose to the chin. The production of each stimulus began and ended in a neutral position, with the mouth closed. Videos (Quicktime movie files, 21 by 21 cm) were displayed centered on a 15-inch MacBook Pro laptop on a black background. Three productions of each /aCa/ stimulus were digitally recorded and audio tracks were equalized in level. Eighteen stimuli (six syllabes × three repetitions) were used to create the AV, AO and VO trials. Stimuli were delivered through Sennheinser HD 121 Pro headphones.

The congruent AV stimuli included digital audio-video files of the speaker saying and articulating the /aCa/ stimuli. For the AO condition, an image of the speaker, appearing neutral and with mouth closed was presented along with the auditory stimulus. For the VO condition, the audio was turned off. Finally, incongruent AV McGurk stimuli were created by carefully combining audio files with non-corresponding video files and matching their onset. We used three repetitions of the two following stimuli: audio /apa/ with visual /aka/ (fusion /ata/) and audio /afa/ with visual /aʃa/ (fusion /asa/).

The total number of items was 180 stimuli (six syllables × three repetitions × three modalities × three types of noise +18 McGurk stimuli, randomly mixed). Four blocks of 45 items were constructed. In each block, the order of appearance of the stimuli was fixed and identical for all participants.

Auditory noise. Each signal was digitalized at a 22,050 Hz sampling frequency. Throughout all conditions of the experiment, stimuli were embedded in noise which was either stationary (i.e., unmodulated), either modulated in amplitude. Modulation in amplitude was achieved by using a white Gaussian noise low-pass filtered at 500 Hz (WGNf). The expression describing the sine-wave modulator, $m(t)$, was

$$m(t) = [1 + \cos(2\pi f_m t)^* WGNf]$$

where the 1st-order modulation frequency f_m was 8 and 128 Hz. The noise was then added to the signal. The SNR was fixed at −23 dB (prior to the 500 Hz filtering). This SNR was determined in a preliminary experiment so as to yield a consonant identification performance of about 40% correct under stationary noise (in AO condition).

Procedure

The experiment was conducted in a dimly-lit quiet room. Participants were seated in front of the laptop and fitted with headphones. Stimuli were presented on a monitor positioned at eye level, 70 cm from the participant's head. Participants were given verbal and written instructions to watch the computer monitor and listen for speech sounds that would be heard over

headphones. They were informed about the identity of the six syllables that would be presented. They were instructed that for some trials, there would be a speech sound but the face would not move (i.e., the AO stimuli), sometimes the face would move but there would be no speech sound (the VO stimuli) and sometimes there is a speech sound and a moving face (i.e., the AV stimuli). No information was given about the presence of the McGurk incongruent stimuli.

Participants were instructed to designate a letter corresponding to the consonant they thought the speaker had said. The six letters were taken from a speech therapist kit (*La planète des Alphas*, Huguenin and Dubois, 2006) which was unknown both from the children with SLI and the children with TLD. Sometimes, children also spontaneously repeated the syllable aloud. Their responses were recorded by the experimenter. They were given 20 practice trials, including AO, VO, and AV congruent stimuli, during which they were provided with feedback regarding the correct responses. Prior to beginning the experimental trials, they were informed that they would no longer receive any feedback.

Following the practice session, participants were presented with the four experimental blocks. The sequence of presentation of the blocks was counterbalanced across participants. After the four experimental blocks, they were given a block of 54 stimuli presented without noise. This quiet block consisted of the six syllables × three repetitions × three modalities (AO, VO, and AV congruent). In a second session, they were submitted to the three language tests.

Participant's percent-correct identification of the VCV syllables presented in each of these conditions served as the dependent measure. For McGurk stimuli, the percent responses corresponding to Audio, Visual and Fusion responses were recorded.

The experiment took place in two 30 min sessions. The first session was devoted to the collection of language measures, and the second one to the experimental data. The experimenter was careful about the attention and concentration of the children, and proposed breaks if necessary.

RESULTS

Results in noise modulated at 8 Hz and noise modulate at 128 Hz were averaged for more clarity and because they were not significantly different.

Single modality conditions

First, results were analyzed in the AO modality in order to ascertain whether our experimental design generated a masking release effect, i.e., higher performances in AM noise than in ST noise. The percentage of correct identification of children with SLI and with TLD for quiet, AM noise, and ST noise, and the masking release effect are presented in **Table 2**. A clear masking release effect was observed for both groups: performance was about 30% better in AM noise than in ST noise.

An ANOVA with repeated measures on Noise (3 levels: quiet, AM, and ST) and Group (children with SLI, children with TLD) was run on these data. The analysis yielded a significant effect of Noise, $[F_{(2,\ 52)} = 327.94,\ p < 0.001]$, and of Group, $[F_{(1,\ 26)} = 4.94,\ p < 0.05]$. The Group × Noise interaction was not significant. Orthogonal contrasts were made on the effect of Noise. The first contrast, comparing the results in quiet on the one hand, and in AM and ST Noise on the other hand, was highly significant, $[F_{(1,\ 26)} = 566.51,\ p < 0.001]$. The second contrast, comparing the results in AM noise and ST noise, was highly significant too, $[F_{(1,\ 26)} = 198.27,\ p < 0.001]$. None of these contrasts interacted with the Group effect. To sum up, performance was better in modulated noise than in stationary noise, and better in quiet than in noisy conditions. The 4.4% difference of masking release between children with SLI and TLD was not significant.

Second, results were analyzed in the VO modality. As expected, children with TDL achieved better performances in VO than children with SLI, regardless of whether the stimuli were presented in quiet, in AM noise, or in ST noise (see **Table 3**). These data were entered in a repeated measures ANOVA, with Group as between subjects factor, and Noise (3 levels: quiet, AM, and ST) as within subjects factor. Only the Group effect was significant, $[F_{(1,\ 26)} = 16.86,\ p < 0.001]$. Neither the Effect of Noise, nor the Group × Noise interactions were significant. To sum up, children with SLI achieved lower performance in identification of syllables presented in speechreading; as expected, auditory noise had no significant effect on the performance in VO.

Congruent AV modality (AV)

Percentages of correct identification of children with SLI and of children with TLD for AV in quiet, in AM noise, and in ST noise are presented in **Table 4**. The performance of children with SLI was significantly lower than the performance of children with TLD in all three conditions.

A repeated measures ANOVA with Noise (3 levels: quiet, AM, and ST) as within-subjects factor and Group (children with SLI, children with TLD) as between-subjects factor was run on these data. The analysis yielded significant effects of Noise, $[F_{(2,\ 52)} = 31.58,\ p < 0.001]$, and of Group, $[F_{(1,\ 26)} = 7.35,\ p < 0.05]$. The Group × Noise interaction was not significant. Orthogonal contrasts were made on the effect of Noise. The first contrast, comparing the results in quiet on the one hand, and in AM and ST

Table 2 | Mean percent correct responses for AO in quiet, AM noise and ST noise, and mean value for the masking release effect.

	SLI	TLD
Silence	97.2 (8.9)	100
AM noise	76.2 (11.4)	85.4 (8.2)
ST noise	47.6 (9.2)	52.3 (10.4)
Masking release	28.6 (10.1)	33.0 (12.9)

Standard deviations are in brackets.

Table 3 | Mean percent correct responses for VO in quiet, AM noise and ST noise.

	SLI	TLD
Silence	54.4 (17.4)	74.5 (7.8)
AM noise	56.7 (12.4)	69.4 (8.0)
ST noise	56.8 (12.4)	71.0 (9.3)

Standard deviations are in brackets.

Table 4 | Mean percent correct responses for AV in quiet, in AM noise, and ST noise, and mean value for Visual Gains (VG).

	SLI	TLD
AV (quiet)	97.6 (3.6)	100
AV/AM	89.1 (10.2)	95.3 (4.7)
AV/ST	83.8 (9.8)	91.3 (6.5)
VG/AM	56.0 (36.7)	61.2 (50.2)
VG/ST	69.1 (17.7)	81.3 (15.3)

Standard deviations are in brackets.

Noise on the other hand, was highly significant, [$F_{(1,\ 26)} = 55.08$, $p < 0.001$]. The second contrast, comparing the results in AM noise and ST noise, was highly significant too, [$F_{(1,\ 26)} = 10.19$, $p < 0.005$]. None of these contrasts interacted with the Group effect.

We calculated the visual gains (VG) in both groups. Visual gain refers to relative increase in AV speech perception performance due to the addition of visual information to the auditory signal (Sumby and Pollack, 1954). We computed VG in ST and AM noise using the following formula:

$$\text{VG/ST} = (\text{AVST} - \text{AOST})/(100 - \text{AOST})$$

$$\text{VG/AM} = (\text{AVAM} - \text{AOAM})/(100 - \text{AOAM})$$

The values of the VG are displayed in **Table 4**. An ANOVA with repeated measures on Noise and Group as between subjects factor yielded no effect of Noise, Group and no interaction.

Overall, the data showed that children with SLI had lower performance on AV syllable identification than children with TLD. However, children with SLI did not differ from children with TLD in masking release effect, nor in visual gain.

McGurk effect

The percentages of auditory, visual and fusion responses were computed relative to the total amount of responses to McGurk stimuli. The distribution of responses is shown in **Figure 1** for children with SLI and TLD children. First, the response pattern of each group was examined to evaluate the impact of noise condition (ST vs. AM) on AV speech integration. Second, the groups were compared in order to examine the effect of language impairment.

In ST noise, children with TLD mainly gave a low rate of auditory responses (5.9%; SD: 10.6), and fusion responses (15.4%; SD: 16.5), and a high rate of visual responses (71.5%; SD: 30.3). Compared to ST noise, children with TLD gave significantly more auditory responses [15.5%; SD: 21.9; $F_{(1,\ 13)} = 6.63$, $p < 0.05$], a higher number of fusion responses [52.4%; SD: 22.6; $F_{(1,\ 13)} = 20.2$, $p = 0.001$], and significantly less visually responses [29.1%; SD: 19.3; $F_{(1,\ 13)} = 39.97$, $p < 0.001$] in AM noise.

In ST noise, children with SLI gave 13.0% (SD: 18.7) of auditory responses, 22.5% (SD: 14.0) of fusion responses, and 48.9% (SD: 23.2) of visual responses. In AM noise, they gave more auditory responses [28.5%; SD: 27.9; $F_{(1,\ 13)} = 13.76$, $p < 0.005$] than in ST noise. The percent of fusion responses (33.3%; SD: 25.6) and of visual responses (32.1%; SD: 23.0) was not significantly different from that in ST noise.

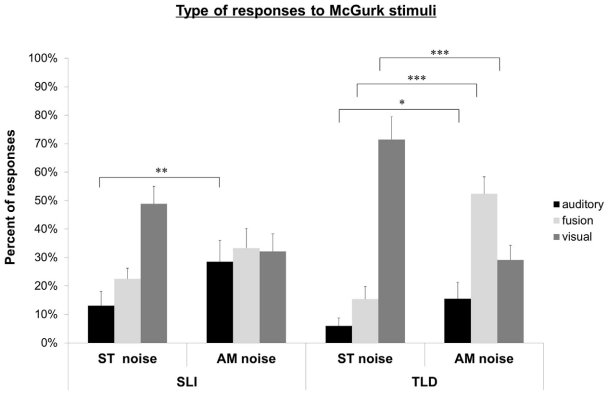


FIGURE 1 | Experiment 1. Auditory, fusion, and visual responses to McGurk stimuli for SLI and TLD groups in ST and AM noise conditions. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Compared to children with TLD, children with SLI had a lower rate of visual responses in ST noise, $F_{(1,\ 26)} = 4.92$, $p < 0.05$, and less fusions in AM noise, $F_{(1,\ 26)} = 4.36$, $p < 0.05$. No other difference was significant.

To sum up, the pattern of responses to McGurk stimuli was clearly modified by the degree of degradation of auditory information in children with TLD: AM noise decreased the rate of visual responses, and increased auditory, and fusion responses. For children with SLI, AM noise increased auditory responses, confirming children's intact auditory masking release effect; however, AM noise has no impact for fusion and visual responses, coherently with SLI's deficit in processing visual speech information.

DISCUSSION

The present study examined the impact of SLI on AV speech perception with a masking release paradigm, already used to study audiovisual integration in TLD children and children with cochlear implants (Huyse et al., 2012). Several results are to be emphasized, in relation to our predictions. First, in AO modality, children with SLI showed a deficit in consonant perception presented in quiet, stationary noise or modulated noise. Despite their speech-in-noise deficit, children with SLI experienced a clear masking release effect, which was not significantly different from that of TLD children: their speech intelligibility was increased in the modulated noise compared to the stationary noise (Ziegler et al., 2005). The average size of the effect was around 30%, which is relatively high compared to the 10% found by Ziegler et al. (2005). Difference between the SNR used in these two studies could be an explanation. Ziegler et al. (2005) used a SNR of 0 dB so as to yield an auditory performance of approximately 50% correct with ST noise. We wanted to obtain a lower level of correct responses in AO/ST, in order to observe both an auditory masking release effect and a visual gain, and we used a SNR of −23 dB. With a lower rate of correct responses in AO/ST as a baseline, it is easier to obtain larger masking release values.

Second, in VO modality, children with SLI were less accurate than TLD children in identification of the six consonants belonging to different visemes. This result is coherent with the notion that children with SLI experience difficulties in perceiving place of articulation (Sussman, 1993; Gerrits and de Bree, 2009), and reveals that this deficit is not specific to auditory processing but could be extended to visual processing (Meronen et al., 2013).

Third, children with SLI performed less well than TLD children in AV congruent modality, indicating that the difficulty of processing of acoustic cues in the AO modality, also impacted audio-visual processing. Surprisingly, the visual gains of children with SLI did not significantly differ from those of the control group.

Fourth, children with TLD were clearly influenced by the degree of degradation of auditory information in AV incongruent modality. In ST noise, when little auditory information is available, participants with TLD mainly relied on visual information. When the speech signal is more available thanks to the existence of noise dips (AM noise), participants with TLD increased their number of auditory responses and their number of fusions even more impressively, while their number of visually-based responses decreased. To sum up, when both auditory and visual information are available (as in AM noise), and participants are able to process them (as are children with TLD), conditions needed to generate McGurk fusions are met. The response pattern of children with SLI to McGurk stimuli was different from that of children with TLD, and coherent with their lower speechreading skills in VO. In ST noise, they gave less visual responses than children with TLD, and in AM noise they reported less fusions than children with TLD. The McGurk effect for the classical pair A/p/V/k/, characterized by a backward shift of the percept from /p/ to /t/ does not work for them to the same degree as for children with TLD. These observations indicate a smaller influence of the visual speech cues on their speech perception processes.

Taken together, the results of Experiment 1 confirm that the difficulties in building accurate phonemic categories is not limited to the auditory modality but is supra-modal in children with SLI. This deficit appears in their responses to stimuli in VO condition, but also to McGurk stimuli. However, the redundancy between visual and auditory information helps children with SLI, as indicated by their visual gain not different from that of children with TLD. Therefore, a more in-depth analysis of how children with SLI process manner, voicing and place of articulation seems necessary in order to get a clearer picture.

Limits of Experiment 1 are the reduced sample of children with language impairment, as well as the number of stimuli used to evaluate the McGurk effect. Therefore, we carried out a second experiment, using a larger set of stimuli. In order to better compare the use of phonetic cues by children with SLI and children with TLD, we also computed the percent of information transmitted for place of articulation, manner, and voicing in AO, AV, and VO.

EXPERIMENT 2

Experiment 2 aimed at generalizing the outcomes of Experiment 1, on a new and larger sample of participants. We introduced several changes in our methodology in order to better evaluate the use of visual information by children with SLI. We included six voiceless and six voiced consonants corresponding to the six visemes used in Experiment 1. Auditory, fusion and visual responses given to McGurk stimuli were measured separately for plosive stimuli A/apa/V/aka/ and A/aba/V/aga/, and fricative stimuli A/afa/V/aʃa/ and A/ava/V/aja/. The interest of the fricative stimuli is that a dominance of the visual responses /ʃ/ or /j/ is observed (Berthommier, 2001; Huyse et al., 2012). If children with SLI recognize /aʃa/ and /aja/ in VO condition, their responses in incongruent AV would show clear influence of visual information, as it is the case in the TLD children. In order to maintain the duration of testing in a reasonable amount of time, only ST noise and AM noise at 8 Hz were used.

In addition to measuring the performance for AO, VO, AV congruent and incongruent stimuli, we computed the specific reception of phonetic features (voicing, place, and manner) by analyses of information transmission (IT) (Miller and Nicely, 1955). Analyses of IT in auditory recognition of speech-in-noise have revealed that children with SLI have a deficit in place, manner, and even more in voicing perception (Ziegler et al., 2005). The present study will allow us to extend these results by examining IT for place, manner and voicing features, in AO, VO, and AV modalities.

We recruited new and larger groups of children with SLI and TDL children whose language performances were examined in a more detailed way (as in Ziegler et al., 2005). We systematically proposed all children to name aloud the syllables in Experiment 2.

MATERIAL AND METHOD

Participants

Fifty-four children, all native and monolingual speakers of French, were recruited as participants. Twenty-seven children (13 boys and 14 girls) constituted the TLD group, and 27 children (17 boys and 10 girls) constituted the group of children with SLI. The two groups were matched as closely as possible by gender, chronological age and by score at the Raven matrices intelligence test (Raven, Court and Raven, 1998). The mean age was 10 years 8 months (range: from 7 years 4 months to 12 years 9 months) for the children with SLI, and 10 years 2 months (from 7 years 6 months to 13 years 8 months) for the TLD children (see **Table 5**). In order to include a child as a participant with SLI, he/she had to present the characteristics outlined in the methodology of Experiment 1.

Hearing and visual abilities of the children with TLD were assessed through a questionnaire filled in by their parents. Children whose parents reported a hearing acuity problem, or who were followed in speech therapy, were removed from the sample.

All children were submitted to the Progressive Matrices Color Raven test (Raven et al., 1998). Language assessment tests included: (a) receptive lexical knowledge (EVIP, French version of the PPVT, Dunn et al., 1993); (b) a standardized test of morpho-syntax, l'E.CO.S.SE (French version of the TROG test, Lecocq, 1996), and (c) Repetition of Difficult Words from the L2MA (Chevrie-Muller et al., 1997). All TLD children presented results comprised between −1.5 SD and +1.5 SD to the three language tests. Reading assessment involved reading aloud Regular

Table 5 | Characteristics of children with SLI and TLD controls in Experiment 2.

	SLI	TLD	Group effect $F_{(1, 26)}$ = *p*-value
Age in years, months (range)	10.9 (7.4–12. 9)	10.2 (7.6–13.9)	*Ns*
Raven (*SD*)	28.44 (4.24)	30.37 (3.56)	$F = 4.06$ $p < 0.05$
EVIP (*SD*)	89.15 (26.91)	116.81 (26.95)	$F = 14.25$ $p < 0.001$
Morpho-syntax (*SD*)	14.29 (6.06)	6.00 (4.24)	$F = 33.94$ $p < 0.001$
Word repetition (*SD*)	15.74 (5.51)	28.11 (1.84)	$F = 122.66$ $p < 0.001$
Irregul. words (*SD*)	8.67 (5.37)	18.19 (1.96)	$F = 74.82$ $p < 0.001$
Regular words (*SD*)	12.03 (5.48)	19.41 (1.05)	$F = 47.11$ $p < 0.001$
Pseudo words (*SD*)	8.15 (4.89)	17.26 (2.03)	$F = 79.76$ $p < 0.001$

Values for Vocabulary indicate raw score on the EVIP, the French version of the PPVT test; Morpho-syntax indicates the number of errors on the ECOSSE picture/sentence word comprehension test. Word repetition values indicate number of correct responses (out of 30) on a sub-test taken from the L2MA language battery. Irregular Words, Regular Words and Pseudo Words values indicate number of correct responses (out of 20) on the reading tests for frequent items taken from Odedys battery.

and Irregular frequent words, and Pseudowords from the battery Odedys-2 (Jacquier-Roux et al., 2005). The characteristics of the participants and a summary of the language test results are found in **Table 5**.

The project has been reviewed and approved by the University research ethic board. Informed consent was obtained from the parents of the participants, and children provided a verbal acceptance prior to their participation. They were informed that they could interrupt their participation if they felt any problem during the experiment.

Stimuli

Movie files of digital AV stimuli were extracted from the same database as those of Experiment 1: /apa/, /afa/, /ata/, /asa/, /aka/, /aʃa/, /aba/, /ava/, /ada/, /aza/, /aga/, and /aʒa/. Three productions of each /aCa/ stimulus were used. The AV (congruent and incongruent), AO, and VO stimuli were constructed in the same way as in Experiment 1. We used four different AV incongruent McGurk stimuli. Two were the classical stimuli with plosive consonants: A/apa/ V/aka/ (→fusion /ata/), and the A/aba/ V/aga/ (→ fusion /ada/). The other two were new combinations based on the fricative pairs: A/afa/ V/aʃa/ (→ fusion /asa/) and A/ava/ V/aʒa/ (fusion /aza/) (Berthommier, 2001). As the recognition of /ʃ/ and /ʒ/ are generally good in speechreading, these fricative pairs offer a new opportunity to examine the processing of visual speech cues by children with SLI.

The AO, VO, and AV stimuli were presented masked by either stationary noise (ST, i.e., unmodulated), or amplitude modulated noise (AM at 8 Hz). The SNR was fixed at −23 dB.

The total amount of items was 252 stimuli (12 syllables × 3 repetitions × 3 modalities × 2 types of noise + 36 McGurk stimuli) randomly mixed and divided in four blocks. In each block, the presentation order of the stimuli was fixed and similar for all participants. In addition, a last bloc containing 120 stimuli (12 syllables × 3 repetitions, × 3 modalities + 12 McGurk stimuli) was presented in quiet, i.e., without noise.

Procedure

The procedure was the same as in Experiment 1, except that participants were instructed to answer by verbally repeating the syllable they perceived. Verbal repetition is an immediate response and is resistant to decay from phonological short-term memory. When the understanding of the syllable was difficult because of articulatory problems, children were encouraged to use a lexical evocation: for example, a child perceiving correctly the syllable /aka/ but pronouncing it /ata/, said I heard /ata/ as in /tamjɔ̃/—the real pronunciation of this word is /kamjɔ̃/ (truck in English). A series of 12 pictures, beginning with the 12 consonants, was prepared to help children to answer. The experimenter recorded the responses. The stimuli in the practice session were representative of the conditions participants would experience in the actual experimental trials, except the McGurk stimuli. For practice trials, subjects were provided with feedback regarding the correct responses. Before beginning the experimental trials, subjects were told that they would no longer receive any feedback.

Following practice, participants were presented with the four experimental blocks. The order of the blocks was counterbalanced across participants. After the four experimental blocks, participants were given a block of 120 stimuli presented in quiet. In a second session, participants were submitted to the Raven matrices, the language and the reading tests.

Participant's percent-correct identification of the syllables presented in each of these conditions served as dependent measure. For McGurk stimuli, we recorded the percent of Auditory, Visual, and Fusion responses.

RESULTS

Single modality conditions

The percentage of correct identification of children with SLI and of the TLD children for stimuli in quiet, AM noise, and ST noise in the AO modality is presented in **Table 6**. Visual inspection of the data revealed that the children with SLI differed from the TLD children in the three conditions. A masking release effect was observed: performance was about 35% better in AM noise than in ST noise for children with SLI, and 39% for children with TLD.

The data were entered in an ANOVA with Noise (quiet, AM, and ST) as within-subjects factor and Group as between-subjects factor. The analysis yielded significant effects of Noise, $F_{(2, 104)} = 2154.92$, $p < 0.001$, and Group, $F_{(1, 52)} = 26.37$, $p < 0.001$. The Noise × Group interaction was just below significance level,

$F_{(1,\ 104)} = 3.01$, $p = 0.054$. The data corresponding to the masking release effect were analyzed with a separate ANOVA, with Group as between-factor: no effect of Group was found ($p = 0.11$). To sum up, children with SLI achieved poorer recognition of auditory speech, but a similar masking release effect as children with TLD.

The percentage of correct identification for stimuli in VO in quiet, AM noise and ST noise is presented in **Table 7**. Children with TLD better identified stimuli in the three conditions than children with SLI. The percent of correct responses for VO stimuli was entered in a repeated measures ANOVA with Noise (quiet, AM noise and ST noise) as within subjects factor, and Group as between subjects factor. The analysis yielded a significant effect of Group, $F_{(1,\ 52)} = 7.69$, $p < 0.01$; and of Noise, $F_{(2,\ 104)} = 5.24$, $p < 0.01$. No interaction was found. To sum up, children with SLI had poorer lipreading performance than children with TLD.

Congruent AV modality (AV)

The percentages of correct identification of children with SLI and of TLD children for AV in quiet, AM noise, and ST noise are presented in **Table 8**. The performance of children with SLI was lower than that of TLD children in the three conditions. The data were entered in a repeated measures ANOVA with Group as between subjects factor and Noise (Quiet, AM Noise, ST Noise) as within subjects factor. The analysis yielded a significant effects of Group, $F_{(1,\ 52)} = 33.41$, $p < 0.001$, and Noise, $F_{(2,\ 104)} = 1.060$, $p < 0.001$. The interaction between Group and Noise was also significant, $F_{(2,\ 104)} = 6.76$, $p < 0.005$. The effect of Noise was further analyzed with two orthogonal contrasts. The first one, comparing the performance in Quiet to the mean performance for AM and ST noise, was highly significant, $F_{(1,\ 52)} = 1589$, $p < 0.001$, as was the interaction with Group, $F_{(1,\ 52)} = 11.95$, $p < 0.005$. The second contrast, comparing performance in AM noise and in ST noise was highly significant, $F_{(1,\ 52)} = 394.05$ $p < 0.001$, and did not interact with Group. To sum up, the effect of Noise on AV speech perception was larger in children with SLI than in children with TLD.

We computed VG/ST and VG/AM noise using the same formula as in Experiment 1. A repeated measures ANOVA with Noise as within subjects factor, and Group as between subjects factor yielded significant effects of Noise, $F_{(1,\ 52)} = 14.46$, $p < 0.001$, and of Group, $F_{(1,\ 52)} = 6.54$, $p < 0.05$. The Group × Noise interaction was not significant. To sum up, children with SLI had lower standardized VG than TLD children, both in ST and AM Noise.

McGurk effects

The percentages of auditory, visual, and fusion responses were computed relative to the total amount of responses to McGurk stimuli. The data have been averaged over the two plosive stimuli, and over the two fricative stimuli.

Plosive McGurk stimuli. The distribution of responses is shown in **Figure 2**. In ST noise, TLD children gave 6.2% (SD: 11.5%) auditory responses, 29.6% (SD: 26.3%) of fusions, and 44.4% (SD: 33.3%) of visual responses. Compared to ST noise, their percent of auditory responses (10.2%, SD: 13.3%) did not change; their percent of fusion responses (43.8%, SD: 27.4%) increased, $F_{(1,\ 26)} = 8.03$, $p < 0.01$; their percent of visual responses (34.9%, SD: 26.5%) significantly decreased, $F_{(1,\ 26)} = 4.2$, $p = 0.05$.

In ST noise, children with SLI gave 6.2% (SD: 12.4%) of auditory responses, 18.5% (SD: 24.6%) of fusions, and 37.6% (SD: 31.6%) of visual responses. Compared to ST noise, their auditory responses increased in AM noise, [12.0%; SD: 13.9%, $F_{(1,\ 26)} = 5.46$, $p < 0.05$], but their rate of fusions (20.4%; SD: 19.2%) and visual responses (35.2%, SD: 24.3%) remained unchanged.

Compared to children with TLD, children with SLI had lower fusion responses in AM noise, $F_{(1,52)} = 13.25$, $p < 0.001$.

Fricative McGurk stimuli. The distribution of responses to fricative McGurk stimuli in the children with SLI group and the children with TLD is shown in **Figure 3**. In ST noise, TLD children gave 1.2% (SD: 4.4%) of auditory responses, 29.0% (SD: 19.4%) of fusions, and 61.7% (SD: 25.2%) of visual responses. Compared to ST noise, TLD children gave a larger number of auditory responses in AM noise [10.5%, SD: 12.1%, $F_{(1,\ 26)} = 11.93$, $p < 0.005$], and a larger number of fusion [68.2%, SD: 23.2%, $F_{(1,\ 26)} = 132.04$; $p < 0.001$]; their rate of visual

Table 6 | Mean percent correct responses for AO in quiet, AM noise and ST noise, and mean values for the masking release effect.

	SLI	TLD
Quiet	94.65 (6.11)	98.87 (1.77)
AM noise	51.65 (9.87)	61.52 (7.17)
ST noise	16.56 (7.76)	22.63 (6.20)
Masking release	35.08 (9.47)	38.89 (7.47)

Standard deviations are in brackets.

Table 7 | Mean percent correct responses for VO in quiet, AM noise and ST noise.

	SLI	TLD
Quiet	28.09 (9.72)	34.88 (9.60)
AM noise	26.85 (8.95)	33.85 (7.90)
ST noise	26.03 (8.69)	30.04 (9.65)

Standard deviations are in brackets.

Table 8 | Mean percent correct responses for AV in quiet, AM noise and ST noise, and mean value for Visual Gains (VG).

	SLI	TLD
AV/Quiet	96.91 (5.57)	99.38 (1.78)
AV/AM	64.40 (7.10)	74.07 (7.22)
AV/ST noise	44.96 (7.35)	52.57 (7.02)
VG/AM	25.33 (13.11)	32.11 (15.98)
VG/ST	33.71 (8.74)	39.92 (9.42)

Standard deviations are in brackets.

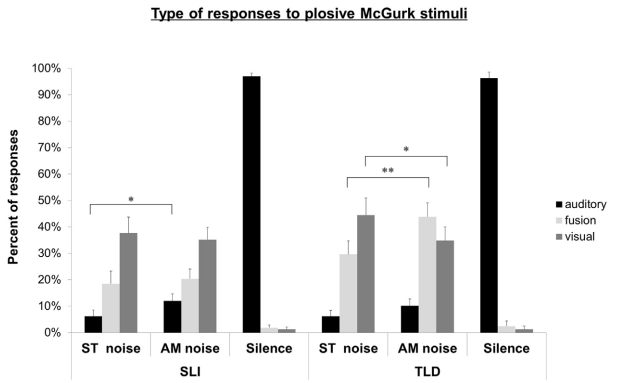


FIGURE 2 | Experiment 2. Auditory, fusion, and visual responses to McGurk plosive stimuli for SLI and TLD groups in ST noise, AM noise and quiet conditions. $^{*}p < 0.05$; $^{**}p < 0.01$.

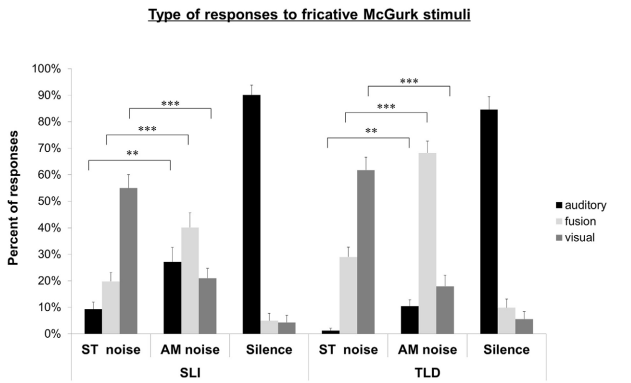


FIGURE 3 | Experiment 2. Auditory, fusion, and visual responses to McGurk fricative stimuli for SLI and TLD groups in ST noise, AM noise and quiet conditions. $^{**}p < 0.01$; $^{***}p < 0.001$.

responses significantly decreased [17.9%, SD: 22.0%, $F_{(1, 26)} = 125.12$; $p < 0.001$].

In ST noise, children with SLI gave 9.3% (SD: 14.1%) of auditory responses, 19.5% (SD: 17.9%) of fusions, and 54.9% of visual responses (SD: 26.5%). Compared to ST noise, they gave significantly more auditory [27.2%; SD: 28.3%; $F_{(1, 26)} = 9.66$; $p < 0.005$], and fusion responses [40.1%; SD: 28.9%; $F_{(1, 26)} = 16.22$; $p < 0.001$]; their rate of visual response significantly decreased [20.9%; SD: 19.25%; $F_{(1, 26)} = 61.37$; $p < 0.001$].

Compared to children with TLD, children with SLI gave more auditory responses in ST and AM noise, $F_{(1, 52)} = 7.93$; $p < 0.01$ and $F_{(1, 52)} = 7.88$, $p < 0.01$ respectively, and less fusions in AM noise, $F_{(1, 52)} = 15.42$, $p < 0.001$. No difference appeared for visual responses.

Finally, in quiet, there was no difference between children with SLI and children with TLD for any kind of response (see **Figures 2**, **3**).

To sum up, the pattern of responses to both plosive and fricative McGurk stimuli was clearly modified by the degree of degradation of auditory information in children with TLD. Compared to ST noise, AM noise decreased the rate of visual responses, and increased auditory and fusion responses. For children with SLI, the pattern was more mixed. AM noise increased the rate of auditory responses for both plosive and fricative, in coherence with their intact auditory masking effect. AM noise increased the rate of fusions, and decreased the rate of visual responses only in the context of fricatives, when the visual information is easily identified. These latter observations are indicative of the audiovisual integration ability of children with SLI.

Phonetic feature information transmission (IT)

The reception of place, manner and voicing features was evaluated by information transmission (IT) analyses performed on the basis of the individual confusion matrices. The percent of IT was averaged over quiet, ST noise and AM noise, and displayed in **Figure 4**.

Because IT was very different in AO or AV than in VO, two separate analyses were run. A repeated measures ANOVA with Feature (place, manner, voicing) and Modality (AO, AV) as within subjects factor, and Group as between subjects factor yielded significant effects of features, $F_{(2, 104)} = 111.41$ $p < 0.001$, of Modality, $F_{(1, 54)} = 925.33$, $p < 0.001$, and of Group, $F_{(1, 52)} = 35.15$, $p < 0.001$. The Modality × Features interaction was significant, $F_{(2, 104)} = 102.22$, $p < 0.001$: IT increases from AO to AV was 17.4% for place, 14.9% for manner, and 8.7% for voicing. No other interaction was significant.

The percent of IT in VO was analyzed with a repeated measures ANOVA with Features (place, manner, voicing) as within subjects factor, and Group as between subjects factor. The analysis yielded a significant effect of Features, $F_{(2, 104)} = 190.86$, $p < 0.001$, and of Group, $F_{(1, 52)} = 11.58$, $p < 0.001$. The Group by Feature interaction was also significant, $F_{(2, 104)} = 9.42$, $p < 0.001$: the difference between groups was large for place of articulation (12.0%), intermediate for manner (6.6%) and almost null for voicing (1.7%).

DISCUSSION

In Experiment 2, new groups of children with SLI and children with TLD matched as closely as possible for gender, chronological age and non-verbal intelligence were tested with an audio-visual masking release paradigm. The identification of syllables in noise was clearly more difficult in Experiment 2 than in Experiment 1. Two reasons may be invoked. A more extensive protocol consisting of voiced and voiceless syllables was administered, and the voiced syllables were more difficult to identify than the voiceless ones. In addition, children with SLI of Experiment 2 could be more language impaired than those of Experiment 1, as indicated by their lower scores on the Word Repetition test.

Despite these differences, the main results of Experiment 2 remarkably replicated the findings of Experiment 1. Children with SLI showed a speech-in-noise deficit, but a masking release effect comparable in size to that of children with TLD. A clear speechreading deficit appeared in children with SLI compared to TLD children. Children with SLI also had less accurate audiovisual speech perception than TLD children. With no surprise, the standardized visual gains of children with SLI were lower than those of TLD children, coherently with the tendency observed

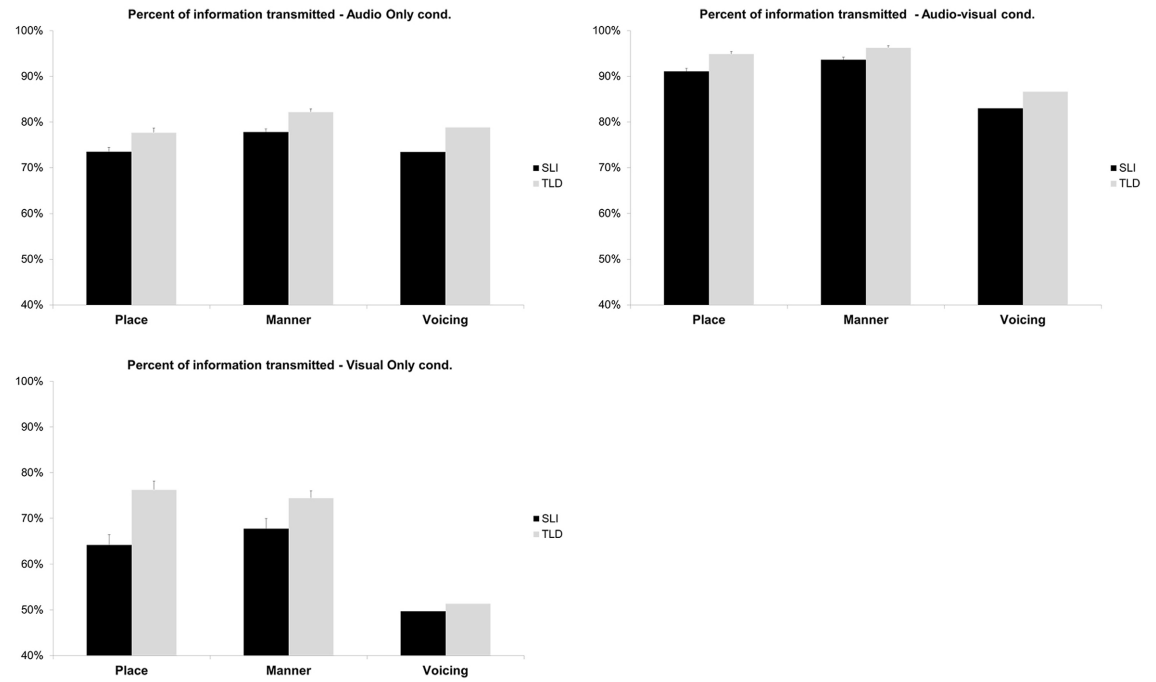


FIGURE 4 | Percent of information transmitted for place, manner, and voicing, as a function of group (SLI vs. TLD) and modality (auditory, audiovisual, and visual).

in Experiment 1 (see **Table 4**). Taken together, the results suggest intact processes at peripheral hearing, but impairment in supra-modal phonemic categorization in children with SLI.

The analysis of percent of information transmitted (IT) is useful to better understand the dynamics of speech perception. A significant increase of percent of IT in AV compared to AO was observed for place of articulation and manner in both groups of children. The complementarity between auditory and visual information is maximal for place of articulation. The fricatives /aʒa/ and /aʃa/, which are highly visible, contribute to the improvement of visual transmission of both manner and place of articulation. Surprisingly, the perception of voicing was improved in AV although voicing was not transmitted by speechreading itself, as indicated by the 50% of IT in the VO modality. It is possible that seeing the movement of the articulators enhances children's attention to the coming sound.

Interestingly, children with SLI showed lower IT percent than children with TLD in the three modalities. This demonstrates that their deficit of IT already found in AO (Ziegler et al., 2005) extends to AV and VO modalities. In VO, the larger difference between the two groups was for place of articulation.

The responses to McGurk stimuli showed an interesting contrast between plosives and fricatives. In the case of plosives A/apa/V/aka/ and A/aba/V/aga/, the visual syllables were poorly identified: 48% by TLD children, and 42% by SLI children. V/aka/ and /aga/ are often confused with V/ata/ and /ada/ (19% in TLD, and 15% in SLI), as well as with V/asa/ and /aza/ (17% in TLD and 19% in SLI). Therefore, the visual information transmitted is reduced and the percent of visual responses is low for both groups. When more auditory information became available, TLD children showed an increase of fusions. SLI children did not but increased their rate of auditory responses.

In the case of fricatives A/afa/V/aʃa/ and A/ava/V/aʒa/, the visual information is easily identified: 84% in TLD and 65% in SLI children. Both groups showed a large amount of visual responses in ST noise (62% for TLD and 55% for SLI children), which significantly decreased in AM noise (18% in TLD and 21% in SLI children). Both groups also showed an increase of auditory and fusion responses in AM, when more auditory was available. The data thus suggest that when children with SLI get access to visual information, they are able to integrate it with auditory information. In other words, the pattern of responses of SLI children is the result of their poorer lipreading skills, but not of a deficit in AV integration. Taken together, the data illustrate the interest of using two types of McGurk stimuli, varying by the degree of availability of the visual information (Berthommier, 2001).

GENERAL DISCUSSION

The aim of the present studies was to test to what extend children with SLI make use of visual articulatory cues to improve their speech-in-noise perception. We used a masking release paradigm, with syllables embedded in ST and AM noise, to which we added the visual information of a talking face (Huyse et al., 2012). Syllables were presented to the participants in three modalities: AO, VO, and AV (congruent and incongruent). We also measured the consonant identification in AO, VO, and AV in quiet. We used child-friendly procedures to elicit the responses to the syllables, i.e., to designate a letter corresponding to the consonant they thought the speaker had said, or to immediately repeat the syllable.

Children with SLI and their age-matched control children performed at, or near, ceiling level when asked to identify syllables in AO or in AV congruent when stimuli were presented in quiet. Our results clearly demonstrate an absence of (or only subtle) difficulty for children with SLI in discriminating /aCa/ syllables under optimal listening conditions. These data confirm previous studies testing AO (Ziegler et al., 2005) or AV speech perception (Norrix et al., 2007; Meronen et al., 2013). Our speech stimuli were produced naturally. Natural speech is rich in redundant acoustic cues and may be easier for children with SLI to perceive than synthetic stimuli (Evans et al., 2002). The good results obtained by children with SLI under optimal conditions also validate our response procedures, which are resistant to decay from phonological short-term memory. Thus, we might be confident that the simple task demands, combined with natural speech, allow us to accurately assess the identification of /aCa/ tokens by SLI and age-control children.

We predicted that children with SLI would experience a speech-in-noise deficit but an intact masking-release effect in the auditory modality (Ziegler et al., 2005). Both expectations were confirmed in Experiments 1 and 2. Contrasting with their good performance under optimal listening conditions, children with SLI show a marked deficit in noisy conditions, confirming their difficulties in separating speech from noise (Sperling et al., 2005; Hornickel et al., 2009; Ziegler et al., 2009). Children with SLI showed a masking release effect of the same size (around 30%) than that of control children. An intact masking release effect is usually taken as a signature of appropriate use of the short temporal minima in the fluctuating background to perceive speech cues, suggesting that the "sensory and cognitive processes known to be involved in masking release, such as auditory grouping based on stimulus spectral and fine-structure cues, perceptual restoration, and informational masking, are functional in children with SLI" (Ziegler et al., 2005, p. 14113). Our data thus support the claim that an intact masking release despite a deficit in speech-in-noise constitutes a robust effect in children with SLI.

If developmental SLI reflects a dysfunction in phonemic categorization as opposed to a purely auditory disorder, we could expect to observe a speechreading deficit. The results of Experiments 1 and 2 clearly showed that children with SLI were less accurate than TLD children in identifying the consonants belonging to six different visemes. Therefore, when speech-in-noise deficit is due to central processing dysfunctioning (rather than to peripherally based auditory problem as in cochlear implantees, see Huyse et al., 2012), the deficit is amodal, and children are less accurate in identifying visual articulatory cues (De Gelder and Vroomen, 1998; Ramirez and Mann, 2005; Norrix et al., 2007; Leybaert and Colin, 2008; Meronen et al., 2013).

Not surprisingly, children with SLI were less influenced by the visual speech cues than TLD children. Clear differences appeared in how participants effectively used visual cues to recover place of articulation when /aCa/ syllables were masked by noise. Children with SLI had lower visual gains both in ST and AM noise (significantly in Experiment 2 and quantitatively in Experiment 1). Again, this result dismisses the idea that the speech perception deficit of children with SLI has a purely auditory basis. Should that be the case, the deficit in the auditory processing domain could be partially circumvented by reliance on visual speech.

The speechreading deficit of SLI children also impacts their response pattern to McGurk stimuli. As expected, TLD children gave mainly visual responses in ST noise, and significantly more auditory and fusions responses in AM noise. In other words, TLD children exhibited a release from masking of the McGurk fusions (Huyse et al., 2012). Children with SLI gave significantly less visual responses than the controls in ST noise (Experiment 1), and less fusions in AM noise (Experiments 1 and 2), confirming previous data (Norrix et al., 2007; Leybaert and Colin, 2008; Meronen et al., 2013).

How to explain the pattern of responses of SLI children to McGurk stimuli ? Do the responses of SLI children result from their lower speechreading skills, or, alternatively, are they the consequence of an atypical integration process itself? On one hand, when visual information is clearly available (as in the fricatives of Experiment 2), children with SLI seem able to integrate auditory and visual information adequately, even if they showed less influence of visual speech. This result is compatible with the "deficit in speechreading skills" hypothesis. On the other hand, it may be that the visual articulatory gestures are processed more independently of the auditory information for children with SLI than for children with TLD. Green (1998) suggested that young children might weight auditory dimensions differently than older children, and alternative weighting might result in reduced interaction with the visual information. Thus, children with SLI may differ from their peers with TLD in terms of how they weight the visual dimensions of the articulated speech segments. We are presently running a new experiment to test more directly these two hypotheses.

The data obtained by children with SLI contrast with those obtained by children with cochlear implant assessed with a similar paradigm (Huyse et al., 2012). In deaf children fitted with a CI, a peripherally based disorder underlies deficits in auditory speech processing in noise; this deficit could be partially circumvented by the introduction of visual articulatory cues. By contrast, a central, amodal deficit in phonemic categorization prevents children with SLI from effectively utilizing these visual articulatory cues. In future, it would be interesting to investigate whether this difference helps identify CI children with SLI.

There are several limitations to the present studies. Children with SLI are a heterogeneous group. It would be interesting to examine whether their speechreading ability and use of visual cues to improve audiovisual speech perception is also variable. Do they differ in linguistic processing of visible articulatory gestures, or do they differ in attentional processes ? Is there a relation between impairment in visible speech processing and potential temporal processing deficits in SLI (see Ten Oever et al., 2013, for a discussion about how AV timing information on articulatory cues aids in syllable identification)?

In addition, the deficit in speech-in-noise perception, poor perception of visual speech, difficulties in fusing auditory and visual stimuli in classic McGurk stimuli could be related to cortical and sub-cortical responses in future studies. According to Hornickel et al. (2009), abnormal encoding of the place of articulation feature of stop consonants should appear in the auditory

brainstem in children with SLI. Such a deficit in the encoding of formant information would lead to representations less resistant to noise, and, possibly, to an under-development of the processing of place of articulation in visual speech, and of integration of auditory and visual speech. In addition, audio-visual integration also has a corresponding cortical response (Colin et al., 2002), which could be absent or reduced in children with SLI.

We can only speculate as to whether or not language training may modify the ability of children with SLI to process the visual speech cues. A long-term study on how speech and language remediation training can help children with SLI more effectively utilize visual articulatory cues in identifying impoverished speech elements may help address this issue better. It would also be interesting to investigate whether their reduced ability to combine auditory and visual information is speech specific, or also occur for other types of integration auditory and visual non-speech, or audio-tactile information. This issue is at the agenda for future research. The outcomes of these types of research will help to better understand the causes of reduced audio-visual speech integration in children with SLI, and to design more adapted rehabilitation programs.

ACKNOWLEDGMENTS

We are very grateful to the Fonds National de la Recherche Scientifique (FNRS, Belgium) for financial support of this project (FRFC Grant n° 2.4539.11). Lucie Macchi is presently at Université de Lille 3. Aurélie Huyse was Aspirant at the FNRS at the time this research took place. Clémence Bayard has a mini-ARC grant from the Université libre de Bruxelles. We are grateful to Benoit Jutras for his comments and discussions on an earlier version of this paper.

REFERENCES

Archibald, L. M. D., and Gathercole, S. E. (2007). Nonword repetition in specific language impairment: more than a phonological short-term memory deficit. *Psychon. Bull. Rev.* 14, 919–924. doi: 10.3758/BF03194122

Berthommier, F. (2001). "Audio-visual recognition of spectrally reduced speech," in *Proceedings of the International Conference on Auditory-Visual Speech Processing* (Aalborg), 183–188.

Bishop, D. V. M. (1998). Development of the Children's Communication Checklist (CCC): a method for assessing qualitative aspects of communicative impairment in children. *J. Child Psychol. Psychiatry* 39, 879–891. doi: 10.1017/S0021963098002832

Bishop, D. V. M., Carlyon, R. P., Deeks, J. M., and Bishop, S. J. (1999). Auditory temporal processing impairment: neither necessary nor sufficient for causing language impairment in children. *J. Speech Lang. Hear. Res.* 42, 1295–1310.

Bishop, D. V. M., and McArthur, G. M. (2004). Immature cortical responses to auditory stimuli in specific language impairment: evidence from ERPs to rapid tone sequences. *Dev. Sci.* 7, F11–F18. doi: 10.1111/j.1467-7687.2004.00356.x

Bishop, D. V. M., and Snowling, M. J. (2004). Developmental dyslexia and specific language impairment: same or different? *Psychol. Bull.* 130, 858–886. doi: 10.1037/0033-2909.130.6.858

Blau, V., Reithler, J., van Atteveldt, N., Seitz, J., Gerretsen, P., Goebel, R., et al. (2010). Deviant processing of letters and speech sounds as proximate cause of reading failure: a functional magnetic resonance imaging study of dyslexic children. *Brain* 133, 868–879. doi: 10.1093/brain/awp308

Bortolini, U., and Leonard, L. B. (2000). Phonology and children with specific language impairment: status of structural constraints in two languages. *J. Commun. Disord.* 33, 131–150. doi: 10.1016/S0021-9924(99)00028-3

Brady, S., Shankweiler, D., and Mann, V. (1983). Speech perception and memory coding in relation to reading ability. *J. Exp. Child Psychol.* 35, 345–367. doi: 10.1016/0022-0965(83)90087-5

Chevrie-Muller, C., Simon, A.-M., and Fournier, S. (1997). *L2MA. Batterie Langage oral et écrit. Mémoire. Attention.* Paris: Editions du Centre de Psychologie Appliquée.

Colin, C., Radeau, M., Soquet, A., Demolin, D., and Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: a phonetic representation within short-term memory. *Clin. Neurophysiol.* 113, 495–506. doi: 10.1016/S1388-2457(02)00024-X

Collet, G., Colin, C., Serniclaes, W., Hoonhorst, I., Markessis, E., Deltenre, P., et al. (2012). Effect of phonological training in French children with SLI: perspectives on voicing identification, discrimination and categorical perception. *Res. Dev. Disabil.* 33, 1805–1818. doi: 10.1016/j.ridd.2012.05.003

De Gelder, B., and Vroomen, J. (1998). Impaired speech perception in poor readers: evidence from hearing and speech reading. *Brain Lang.* 64, 269–281. doi: 10.1006/brln.1998.1973

Dunn, L. M., Thériault-Whalen, C., and Dunn, L. (1993). *Echelle de vocabulaire en images Peabody. Adaptation Française du Peabody Picture Vocabulary Test-revised.* Toronto: Psycan.

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *J. Speech Hear. Res.* 15, 413–422.

Evans, J. L., Viele, K., Kass, R. E., and Tang, F. (2002). Grammatical morphology and perception of synthetic and natural speech in children with specific language impairments. *J. Speech. Lang. Hear. Res.* 45, 494–504. doi: 10.1044/1092-4388(2002/039)

Ferguson, M. A., Hall, R. L., Riley, A., and Moore, D. R. (2011). Communication, listening, cognitive and speech perception skills in children with auditory processing disorder (APD) or specific language impairment (SLI). *J. Speech Lang. Hear. Res.* 54, 211–227. doi: 10.1044/1092-4388(2010/09-0167)

Füllgrabe, C., Berthommier, F., and Lorenzi, C. (2006). Masking release for consonant features in temporally fluctuating background noise. *Hear. Res.* 211, 74–84. doi: 10.1016/j.heares.2005.09.001

Gerrits, E., and de Bree, E. (2009). Early language development of children at familial risk of dyslexia: speech perception and production. *J. Commun. Disord.* 42, 180–194. doi: 10.1016/j.jcomdis.2008.10.004

Grant, K. W., and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208. doi: 10.1121/1.1288668

Green, K. P. (1998). "The use of auditory and visual information during phonetic processing: implications for theories of speech perception," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, eds R. Campbell, B. Dodd, and D. Burnham (Hove: Psychology Press), 3–25.

Hornickel, J., Skoe, E., Nicol, T., Zecker, S., and Kraus, N. (2009). Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13022–13027. doi: 10.1073/pnas.0901123106

Huguenin, C., and Dubois, O. (2006). *Méthode Alpha Ludique et Efficace Pour le Déclic Lecture.* Chalon-sur-Saône: Eveil et découvertes.

Huyse, A., Berthommier, F., and Leybaert, J. (2012). Degradation of labial information modifies audiovisual speech perception in cochlear implanted children. *Ear Hear.* 34, 110–121. doi: 10.1097/AUD.0b013e3182670993

Jacquier-Roux, M., Valdois, S., and Zorman, M. (2005). *Odédys-2. Outil de Dépistage des Dyslexies.* Grenoble: Laboratoire Cogni-sciences - IUFM.

Johnson, E. P., Pennington, B. F., Lee, N. R., and Boada, R. (2009). Directional effects between rapid auditory processing and phonological awareness in children. *J. Child Psychol. Psychiatry* 50, 902–910. doi: 10.1111/j.1469-7610.2009.02064.x

Lecocq, P. (1996). *L'E.CO.S.SE. Une Epreuve de Compréhension Syntaxico-Semantique*. Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Leonard, L. B. (1998). *Children with Specific Language Impairment*. Cambridge, MA: MIT Press.

Leonard, L. B. (2004). "Specific language impairment in children," in *The MIT Encyclopedia of Communication Disorders*, ed R. Kent (Cambridge, MA: MIT Press), 402–405.

Leonard, L. B. (2009). Some reflections on the study of children with specific language impairment. *Child Lang. Teach. Ther.* 25, 169–171. doi: 10.1177/0265659009105891

Leybaert, J., and Colin, C. (2008). "Perception multimodale de la parole dans le développement normal et atypique: premières données," in *Apprentissage des Langues*, eds M. Kail, M. Fayol, and M. Hickman (Paris: CNRS Editions), 529–547.

MacLeod, A., and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *Br. J. Audiol.* 21, 131–141. doi: 10.3109/03005368709077786

Maillart, C., and Parisse, C. (2006). Phonological deficits in French speaking children with SLI. *Int. J. Lang. Commun. Disord.* 41, 253–274. doi: 10.1080/13682820500221667

McArthur, G. M., and Bishop, D. V. M. (2005). Speech and non-speech processing in people with specific language impairment: a behavioral and electrophysiological study. *Brain Lang.* 94, 260–273. doi: 10.1016/j.bandl.2005.01.002

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Meronen, A., Tiippana, K., Westerholm, J., and Ahonen, T. (2013). Audiovisual speech perception in children with developemental language disorder in degraded listening conditions. *J. Speech Lang. Hear. Res.* 56, 211–221. doi: 10.1044/1092-4388(2012/11-0270)

Miller, G. A., and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 338–352. doi: 10.1121/1.1907526

Mody, M., Studdert-Kennedy, M., and Brady, S. (1997). Speech perception deficits in poor readers: auditory processing or phonological coding? *J. Exp. Child Psychol.* 64, 199–231. doi: 10.1006/jecp.1996.2343

Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A. (2003). Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. *J. Acoust. Soc. Am.* 113, 961–968. doi: 10.1121/1.1531983

Nittrouer, S., Shune, S., and Lowenstein, J. H. (2011). What is the deficit in phonological processing deficits: auditory sensitivity, masking, or category formation? *J. Exp. Child Psychol.* 108, 762–785. doi: 10.1016/j.jecp.2010.10.012

Norrix, L. W., Plante, E., Vance, R., and Boliek, C. A. (2007). Auditory-visual integration for speech by children with and without specific language impairment. *J. Speech Lang. Hear. Res.* 50, 1639–1651. doi: 10.1044/1092-4388(2007/111)

Ramirez, J., and Mann, V. (2005). Using auditory-visual speech to probe the basis of noise-impaired consonant-vowel perception in dyslexia and auditory neuropathy. *J. Acoust. Soc. Am.* 118, 1122–1133. doi: 10.1121/1.1940509

Raven, J. C., Court, J. H., and Raven, J. (1998). *Progressive Matrices Couleur*. Paris: Editions du Centre de Psychologie Appliquée.

Robertson, E. K., Joanisse, M. F., Desroches, A. S., and Ng, S. (2009). Categorical speech perception deficits distinguish language and reading impairments in children. *Dev. Sci.* 12, 753–767. doi: 10.1111/j.1467-7687.2009.00806.x

Rosen, S. (2003). Auditory processing in dyslexia and specific language impairment: is there a deficit? What is its nature? Does it explain anything? *J. Phon.* 31, 509–527. doi: 10.1016/S0095-4470(03)00046-9

Rosen, S., Adlard, A., and Van der Lely, H. K. J. (2009). Backward and simultaneous masking in children with grammatical specific language impairment: no simple link between auditory and language abilities. *J. Speech Lang. Hear. Res.* 52, 396–411. doi: 10.1044/1092-4388(2009/08-0114)

Sperling, A. J., Lu, Z. L., Manis, F. R., and Seidenberg, M. S. (2005). Deficits in perceptual noise exclusion in developmental dyslexia. *Nat. Neurosci.* 8, 862–863. doi: 10.1038/nn1474

Studdert-Kennedy, M., and Mody, M. (1995). Auditory temporal perception deficits in the reading-impaired: a critical review of the evidence. *Psychon. Bull. Rev.* 2, 508–514. doi: 10.3758/BF03210986

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Summerfield, Q. (1987). "Some preliminaries to a theory of audiovisual speech processing," in *Hearing by Eye II: The Psychology of Speechreading and Audiovisual Speech Processing*, eds R. Campbell, B. Dodd, and D. Burnham (Hove: Erlbaum Associates), 58–82.

Sussman, J. E. (1993). Auditory processing in children's speech perception: results of selective adaptation and discrimination tasks. *J. Speech Hear. Res.* 36, 380–395.

Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. *Brain Lang.* 9, 182–198. doi: 10.1016/0093-934X(80)90139-X

Tallal, P., Miller, S. L., and Fitch, R. (1993). Neurobiological basis of speech: a case for the preeminence of temporal processing. *Ann. N.Y. Acad. Sci.* 682, 27–47. doi: 10.1111/j.1749-6632.1993.tb22957.x

Tallal, P., and Piercy, M. (1973). Defects in nonverbal auditory perception in children with developmental aphasia. *Nature* 241, 468–469. doi: 10.1038/241468a0

Ten Oever, S., Sack, A., Wheat, K. L., Bien, N., and van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Front. Psychol.* 4:331. doi: 10.3389/fpsyg.2013.00331

Wechsler, D. (1996). *WISC-III. Echelle d'intelligence de Wechsler pour enfants, Troisième édition.* Paris: Editions du Centre de Psychologie Appliquée.

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M., and Merzenich, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature* 387, 176–178. doi: 10.1038/387176a0

Ziegler, J. C., Pech-Georgel, C., George, F., Alario, F.-X., and Lorenzi, C. (2005). Deficits in speech perception predict language learning impairment. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14110–14115. doi: 10.1073/pnas.0504446102

Ziegler, J. C., Pech-Georgel, C., George, F., and Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Dev. Sci.* 12, 732–745. doi: 10.1111/j.1467-7687.2009.00817.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 February 2014; paper pending published: 21 March 2014; accepted: 22 April 2014; published online: 20 May 2014.

Citation: Leybaert J, Macchi L, Huyse A, Champoux F, Bayard C, Colin C and Berthommier F (2014) Atypical audio-visual speech perception and McGurk effects in children with specific language impairment. Front. Psychol. ***5****:422. doi: 10.3389/fpsyg.2014.00422*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 July 2014
doi: 10.3389/fpsyg.2014.00754

Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex

Jack C. Rogers[1,2] *, Riikka Möttönen[1], Rowan Boyles[1] and Kate E. Watkins[1]

[1] *Department of Experimental Psychology, University of Oxford, Oxford, UK*
[2] *School of Psychology, University of Birmingham, Birmingham, UK*

Edited by:
Jean-Luc Schwartz, Centre National de la Recherche Scientifique, France

Reviewed by:
Ivan Toni, Radboud University, Netherlands
Marc Sato, Centre National de la Recherche Scientifique and Grenoble University, France

****Correspondence:***
Jack C. Rogers, School of Psychology, University of Birmingham, Edgbaston, B15 2TT Birmingham, UK
e-mail: j.rogers@bham.ac.uk

Perceiving speech engages parts of the motor system involved in speech production. The role of the motor cortex in speech perception has been demonstrated using low-frequency repetitive transcranial magnetic stimulation (rTMS) to suppress motor excitability in the lip representation and disrupt discrimination of lip-articulated speech sounds (Möttönen and Watkins, 2009). Another form of rTMS, continuous theta-burst stimulation (cTBS), can produce longer-lasting disruptive effects following a brief train of stimulation. We investigated the effects of cTBS on motor excitability and discrimination of speech and non-speech sounds. cTBS was applied for 40 s over either the hand or the lip representation of motor cortex. Motor-evoked potentials recorded from the lip and hand muscles in response to single pulses of TMS revealed no measurable change in motor excitability due to cTBS. This failure to replicate previous findings may reflect the unreliability of measurements of motor excitability related to inter-individual variability. We also measured the effects of cTBS on a listener's ability to discriminate: (1) lip-articulated speech sounds from sounds not articulated by the lips ("ba" vs. "da"); (2) two speech sounds not articulated by the lips ("ga" vs. "da"); and (3) non-speech sounds produced by the hands ("claps" vs. "clicks"). Discrimination of lip-articulated speech sounds was impaired between 20 and 35 min after cTBS over the lip motor representation. Specifically, discrimination of across-category ba–da sounds presented with an 800-ms inter-stimulus interval was reduced to chance level performance. This effect was absent for speech sounds that do not require the lips for articulation and non-speech sounds. Stimulation over the hand motor representation did not affect discrimination of speech or non-speech sounds. These findings show that stimulation of the lip motor representation disrupts discrimination of speech sounds in an articulatory feature-specific way.

Keywords: continuous theta-burst stimulation (cTBS), transcranial magnetic stimulation (TMS), primary motor cortex, auditory discrimination, sensorimotor, categorical perception

INTRODUCTION

Our ability to categorize acoustic speech signals is integral to accurate speech perception. Rather than perceiving continuous variations in speech in a linear fashion, variations along an acoustic continuum tend to be perceived categorically. A hallmark of categorical perception is that listeners are better at discriminating two sounds from opposite sides of the phonetic category boundary compared to two sounds with an equivalent acoustic distance that fall on the same side of the category boundary (Liberman et al., 1957; Repp, 1984). According to Liberman's motor theory of speech perception (Liberman et al., 1967; Liberman and Mattingly, 1985) the listener perceives speech by simulating the "intended articulatory gestures" of the speaker and this affects the ability to categorize speech sounds.

This proposed link between speech perception and production remains a topic of active investigation and debate (e.g., Scott et al., 2009; Pulvermüller and Fadiga, 2010; Hickok et al., 2011). A series of studies have shown that listening to speech activates parts of the premotor and primary motor (M1) cortex in the brain that are important for speech production (e.g., Fadiga et al., 2002; Watkins et al., 2003; Wilson et al., 2004; Pulvermüller et al., 2006; Roy et al., 2008; Murakami et al., 2011). Functional imaging activity is observed in the lip and tongue representations in M1 during listening to speech sounds produced using the lip and tongue articulators (e.g., the phonemes /p/ and /t/), respectively (Pulvermüller et al., 2006). Studies using single pulses of transcranial magnetic stimulation (TMS) over the lip representation of the left M1 to elicit motor-evoked potentials (MEPs) in the lip muscles found that listening to speech enhanced motor excitability (Watkins et al., 2003; Murakami et al., 2011). Similarly, MEPs recorded from the tongue in response to single-pulse TMS showed facilitation specifically when participants listened to words that included speech sounds produced by the tongue (Fadiga et al., 2002).

Despite this growing body of evidence, the functional role of motor representations of articulators in speech perception remains unclear. Brain imaging and single-pulse TMS studies that demonstrate increased activity or excitability of motor areas during speech perception cannot answer key questions about whether these changes contribute to speech perception or are

merely correlates of it. It is possible to examine whether these regions contribute to speech perception by using repetitive TMS (rTMS) to temporarily disrupt activity in the motor cortex. Interfering with the function of a specific cortical area (i.e., using TMS to create a "virtual lesion") allows exploration of causal relationships between the stimulated brain region and behavioral performance (see Devlin and Watkins, 2007; Möttönen et al., 2014). Several previous studies using such methods demonstrated the contribution of the left premotor or primary motor cortex (M1) to performance in speech perception tasks (e.g., Meister et al., 2007; D'Ausilio et al., 2009; Sato et al., 2009; Bartoli et al., 2013). For instance, low-frequency rTMS has been shown to suppress motor excitability of the lip representation in left M1 temporarily (e.g., Möttönen and Watkins, 2009). This TMS-induced disruption of the motor lip representation also impaired the ability of listeners to categorically perceive and discriminate speech sounds drawn from acoustic continua ranging between lip- and tongue-articulated phonemes (e.g., "ba" vs. "da" and "pa" vs. "ta"; Möttönen and Watkins, 2009). The disruption did not impair the ability to categorically perceive or discriminate sounds from acoustic continua that are not articulated by the lips (e.g., "ka" vs. "ga" and "da" vs. "ga"). The effect was also specific to the site of stimulation, since disruption of the hand representation within left M1 had no effect on behavioral performance. These findings suggest that the motor representation of the articulators in left M1 contributes to discrimination of speech sounds in an articulator-specific way.

One methodological limitation of low-frequency rTMS, however, is that the duration of the observed modulatory effect is roughly equivalent to the length of stimulation (i.e., the effects last approximately 15 min following 15 min of rTMS). Another form of rTMS, continuous theta-burst stimulation (cTBS), has been shown to produce long-lasting (e.g., 60 min) suppression of motor excitability following only a short train (e.g., 40 s) of stimulation with maximum effects occurring between 20 and 40 min after cTBS (Huang et al., 2005). During cTBS, low-intensity bursts of high-frequency (50 Hz) rTMS are repeated at 5 Hz (i.e., the theta-frequency). Even though adverse effects attributed to theta-burst stimulation (TBS) are reported to be extremely mild and infrequent (e.g., Grossheinrich et al., 2009; Oberman et al., 2011), safety guidelines regarding the use of TBS have yet to be published. A degree of caution in its application is advised, therefore. Here, we used cTBS to stimulate the motor representation of the lips in M1, which allowed us a longer window of time during which we could test auditory discrimination abilities for a wider range of stimuli than tested in previous studies.

In the current study, we delivered 40 s of cTBS over the lip or hand representation of left M1. We assessed changes in cortical excitability within each target region over time by recording MEPs from the lips and hand before and after cTBS. The main aim of the experiment was to replicate and extend our previous findings using low frequency rTMS (Möttönen and Watkins, 2009), by assessing whether cTBS over the lip representation in M1 also impairs discrimination of speech sounds that require the lips for articulation. It has been suggested that rTMS-induced impairments in behavioral performance observed previously in the context of a same-different paradigm (e.g., Möttönen and Watkins, 2009) may reflect changes in response bias rather than perceptual processes important for speech (Hickok, 2010). A potential disadvantage of the same–different paradigm is that listeners may favor one of the response alternatives, resulting in a subjective bias towards "same" or "different" responses (Gerrits and Schouten, 2004; Macmillan and Creelman, 2005). We aimed to avoid this potential confound by using a variant of the ABX-discrimination task, AXB, the prototypical discrimination test used for assessing categorical perception. In an AXB-type task, the second stimulus (X) is identical to either the first (A) or the third (B) stimulus. All stimuli in this study were presented at two different inter-stimulus intervals (ISIs; 200 and 800 ms). Previous studies have shown that variations are retained in acoustic short-term memory if a short ISI (200–300 ms) is used (Pisoni and Lazarus, 1974; Pisoni and Tash, 1974; Pisoni, 1977; Massaro and Cohen, 1983). If the ISI exceeds the life span of auditory memory then an abstract, phonetic label based on pre-established categories is used to discriminate speech sounds (Massaro and Cohen, 1983; Gerrits and Schouten, 2004). Manipulating the ISI between sounds provided an opportunity to assess potential differences in discrimination strategy related to auditory memory versus phonetic-categorization. An impairment in discrimination resulting from TMS over the lip representation in M1, particularly at the longer ISI (800 ms), would also be consistent with findings from previous TMS studies demonstrating a role for the motor system in phonological segmentation and verbal working memory processes (Romero et al., 2006; Sato et al., 2009). The current study also differed further from our previous work in that the stimuli included recordings of natural speech sounds from which high-quality place-of-articulation continua were generated using a channel-vocoder ("Straight"; Kawahara et al., 1999). The continua ranged from lip- to tongue-articulated phonemes ("ba"–"da") and phonemes that do not involve the lips in their articulation ("da"–"ga"). In addition to speech sounds, we also aimed to determine whether cTBS-induced disruption of the hand motor representation affected discrimination of non-speech sounds produced by the hands. The non-speech stimuli comprised auditory continua ranging from "clap" sounds (both hands clapped together) to "click" sounds (generated by striking the thumb on the middle finger).

The main aim of the current study was to further investigate the specificity of TMS-induced motor disruptions on auditory discrimination. We predicted that cTBS over the lip representation of M1 would impair discrimination of lip-articulated speech sounds (i.e., "ba" vs. "da") but not of sounds that did not require the lips in their production. We also predicted that disruption of the lip motor representation would not affect discrimination of the non-speech control sounds produced by the hands. However, we anticipated a possible double-dissociation whereby cTBS over the hand motor representation would impair discrimination of non-speech but not speech sounds.

MATERIALS AND METHODS

PROCEDURE

Continuous TBS was applied over the left primary motor cortex at the level of either the lip or the hand representation. We assessed the behavioral effects of cTBS on the ability of participants to

discriminate speech and non-speech sounds. The sound stimuli were drawn from three acoustic continua ranging between lip- and tongue-articulated phonemes ("ba"–"da"), another three continua created from phonemes that do not involve the lips in their articulation ("ga"–"da") and three non-speech continua created from recordings of sounds made by the hands ("clap"–"click").

Participants attended two testing sessions on separate days. During the first session, an identification task was carried out (see below for details). This allowed us to determine subject-specific logistic curves and category boundaries for the test continua prior to the second session. During the first session participants were also familiarized with the TMS equipment and set-up. This ensured that MEPs could be measured in both the contracted lip and hand muscles using single-pulse TMS. In the second session, participants performed an AXB-type discrimination task on the sound stimuli before and after receiving cTBS to either the hand or the lip representation. MEPs from the target muscle (lip or hand) were elicited using single pulse TMS to assess the effect of TBS on motor excitability (**Figure 1**).

PARTICIPANTS

Twenty-seven right-handed native English speakers participated in this experiment. One participant withdrew due to discomfort during testing. Two participants did not complete the second session because they did not show categorical perception of either the speech or the non-speech sound continua in the first session. Data obtained from a fourth participant were excluded because the MEPs recorded during the second session were unreliable indicating a problem with the coil placement. Data from the remaining twenty-three participants were analyzed; hand stimulation group ($n = 11$, 18–45 years, 5 female), lip stimulation group ($n = 12$, 18–45 years, 5 female). All participants were medication-free with no personal or family history of seizures or other neurological disorders. All had normal or corrected-to-normal vision and reported no hearing problems. Informed consent was obtained from each participant before the experiment. All experiments were performed under permission from Oxfordshire NHS Research Ethics Committee B (REC Reference Number 10/H0605/7).

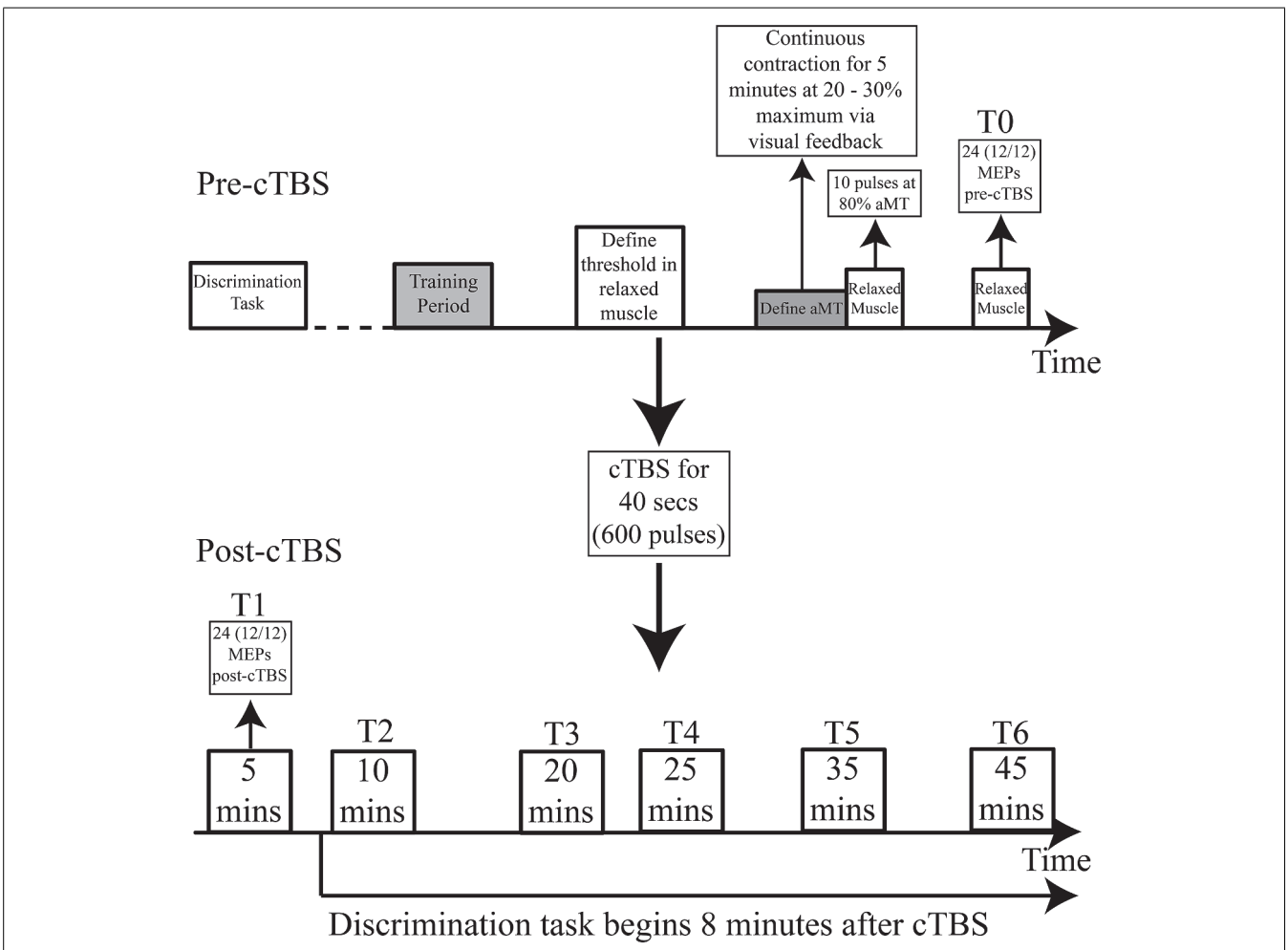


FIGURE 1 | Experimental procedure: cTBS applied for 40 s. 12 MEPs recorded from the target (lip or hand) and non-target (lip or hand) muscle (24 in total) before cTBS (T0) and 5 (T1), 10 (T2), 20 (T3), 25 (T4), 35 (T5), and 45 (T6) min after cTBS. The pre-cTBS discrimination task was completed before TMS set-up began. The post-cTBS discrimination task began 8 min after theta-burst stimulation. aMT = active motor threshold.

BEHAVIORAL TASKS

Behavioral tasks were controlled and presented using Presentation software (Neurobehavioural systems) with all stimuli delivered through insert earphones (Etymotic Research). The earphones also served to protect the participant's hearing during TMS. Participants were familiarized with all tasks and stimuli before testing.

Stimuli

Time-aligned averaging of periodic, aperiodic and F0 representations in the "Straight" channel-vocoder (Kawahara et al., 1999) was used to generate 10-step audio-morphed continua between pairs of naturally recorded speech (/ba/–/da/ and /ga/–/da/ speech syllables) and non-speech (/clap/–/click/) sounds (eight 10-step speech continua and four 10-step non-speech continua). To ensure that equivalent positions in the pairs of sounds were averaged, dynamic-time warping (www.ee.columbia.edu/∼dpwe/resources/matlab/) was used implemented in Matlab (The Mathworks, Natick, MA, USA). This ensured that anchor points placed at evenly spaced positions in sound token 1 (at 50-ms intervals) could then be mapped onto a maximally similar corresponding position in sound token 2. This provides an automated means of creating high-quality, natural-sounding continua and allows us to use the proportion of sound token 1 compared to sound token 2 as a dependent measure when combining responses to different continua. For instance, for each pair, we generated 10 intermediate tokens as 10% acoustic steps from 5% (highly similar to sound 1, e.g., "ba" or "ga" or a "clap" sound) through to 95% (highly similar to sound 2, e.g., "da" or a "click" sound). A 45 or 55% sound token is likely to be heard as perceptually ambiguous, and may, for example, be interpreted as "ba" or "da" depending on the listener and the context.

Pilot experiment

The final stimuli were three "ba"–"da" continua produced by two male speakers and one female speaker, three "ga"–"da" continua spoken by one male speaker and two female speakers and three non-speech "clap"–"click" continua. These continua were chosen based on identification responses and category boundary values obtained from a pilot identification task. Sixteen native-English speaking participants (none of whom subsequently participated in the experiment described here) heard each of the 120 generated sound tokens (10 tokens for each of the 12 sound continua; eight 10-step speech continua and four 10-step non-speech continua) five times (600 tokens altogether split evenly across four blocks). They were then provided with a visual prompt 500 ms after stimulus offset highlighting two possible alternatives (e.g., "ba" or "da") and responded with a key-press to indicate which of the two-alternatives they heard. A third-response alternative was also offered if participants believed they heard something other than the two-alternatives presented on the screen. Proportions of responses for each token were averaged over participants and transformed such that a logistic function could be fitted to the data for each pair and the position of the category boundary (i.e., the estimated morphing percentage for which equal numbers of sound token 1 and 2 responses might be expected) could be computed. Selecting the stimulus continua in this way ensured that (1) the category boundary was close to 50% for both speech and non-speech sounds and (2) that there was no significant difference in boundary position across the different stimulus continua [$F(2,30) = 1.41$, $p > 0.1$]. Analysis of occasions when listeners reported hearing something other than the two-alternatives presented on the screen revealed an average of 2.75% "other" responses (range 0.88–5.13%) for the nine chosen stimulus continua, ($F < 1$).

Identification task

In the identification task (first session), participants were presented with 12 repetitions of the 10 sound tokens from each of the nine continua (1080 trials in total split across four blocks) in a pseudo-randomized sequence. Participants saw a prompt 500 ms after stimulus offset indicating two possible alternatives ("ba"–"da", "ga"–"da" or "clap"–"click"; left or right side stimulus presentation counterbalanced across trials) and responded with a key-press to indicate which of the two-alternatives they heard. Order of presentation of trials from each stimulus pair was pseudo-randomized across each of the blocks using MIX software (van Casteren and Davis, 2006). This ensured that no more than three exemplars from one stimulus pair (i.e., "clap"–"click", "ba"–"da" or "ga"–"da"), no more than two exemplars from the same speaker and no more than two exemplars from the same point along the continuum were heard in succession. Interrogation of the subject-specific responses obtained during the identification task ensured that the category boundary position was between 35 and 65% for all participants and for all speech and non-speech continua.

Analysis of identification data

Logistic regression was used to fit curves to each participant's identification data and obtain slopes (gradient) and the position of the category boundary for each acoustic continuum (i.e., "ba"–"da", "ga"–"da" and "clap"–"click"). These were computed using the formula:

$$y = \frac{e^{\beta_0+\beta 1}}{1 + e^{\beta_0+\beta 1}}$$

where e is the exponent function and $\beta_0 + \beta_1$ refers to the regression line, with β_0 representing the constant and β_1 representing the gradient/regression coefficient. The higher the value of β_1 the steeper the logistic curve (i.e., category boundary). By calculating the parameters of β_0 (constant) and β_1 (gradient) for this fitted function it is also possible to compute the position of the category boundary along the blended stimulus continuum, which corresponds to 50% accuracy on the y-axis.

Discrimination task

In the discrimination task (second session), participants heard triplets of sounds one of which differed from the other two by three steps (i.e., 30%) along the acoustic continuum from which the sounds were drawn. The 30% change along the continuum could be either within-category (5% vs. 35% or 65% vs. 95%) or across-category (35% vs. 65%). Participants were required to indicate as accurately as possible with a left-hand button press whether the first (A) or the last sound (B) was different to the middle sound (X) in a triplet. An equal number of AAB and ABB type trials were

presented with the stimulus pairs presented in both a forward (e.g., "ba"–"ba"–"da", "ba"–"da"–"da") and backward (e.g., "da"–"da"–"ba", "da"–"ba"–"ba") direction. This ensured that the position of the "different" sound in the triplet was not predictable. Thus there were three triplets (two within category and one across category) for each of the nine generated continua (three continua per contrast; three contrasts) that were presented as AAB or ABB, forwards and backwards ($3 \times 9 \times 2 \times 2 = 108$ triplets). Each triplet was repeated three times with an inter-stimulus interval (ISI) between sounds in each triplet that was either 200 or 800 ms (six times in total; 648 stimuli; see **Figure 2**).

The AXB-type discrimination task was performed before and after cTBS. Before cTBS, the task was split into three blocks between which participants rested. Participants also received a 15-min break after completion of the discrimination task and prior to receiving cTBS. The post-rTMS discrimination task began 8-min after cTBS. At fixed time-points after cTBS, visual cues appeared on screen alerting the participant to "STOP. Take a break." And the experimenter was cued to "Apply TMS now." Single pulses of TMS were applied during the breaks over both the hand and the lip representation to elicit MEPs in the target and non-target muscles. All participants completed the discrimination task prior to the 45-min time-point at which the final set of 24 MEPs was recorded or by 50 min (shortly after the final set of 24 MEPs).

ELECTROMYOGRAPHY (EMG)

Electromyography (EMG) activity was recorded from the lip and hand muscle site via surface electrodes (22 mm × 30 mm Kendall® ABRO neonatal electrocardiogram electrodes). The electrodes were attached to the right corners of the upper and lower lip (orbicularis oris) and to the first dorsal interosseous (FDI) muscle and index finger of the right hand. The ground electrode was attached to the right temple in all cases. EMG signals were amplified, band-pass filtered (0.1–1000 Hz) and sampled (5000 Hz) via a CED 1902 four-channel amplifier, a CED 1401 analog-to-digital converter and a PC running Spike2 software (version 7; Cambridge Electronic Design).

All participants completed an initial "training period" during which they were required to produce a constant level of contraction of the lip or hand muscles via visual feedback indicating the level of online EMG activity displayed as power spectra. This training period continued for approximately 5 min or until a satisfactory level of 20–30% maximum voluntary contraction was reached as determined by two experimenters. Participants produced this level of contraction whilst single pulses of TMS were applied over the cortex to determine the active motor threshold (aMT) at the hand or lip representation "hot spot" (Möttönen et al., 2014).

TMS

Magnetic stimulation was given over the hand or lip area of motor cortex and delivered using a hand-held 70 mm figure-eight coil (Magstim Co., Whitland, Camarthenshire, UK). Monophasic single pulses were generated by a Magstim 200 stimulator and used to elicit MEPs. Biphasic pulses were generated by a Magstim Super Rapid^2 and used to define the aMT and deliver cTBS. The coil was placed tangentially to the scalp with the induced

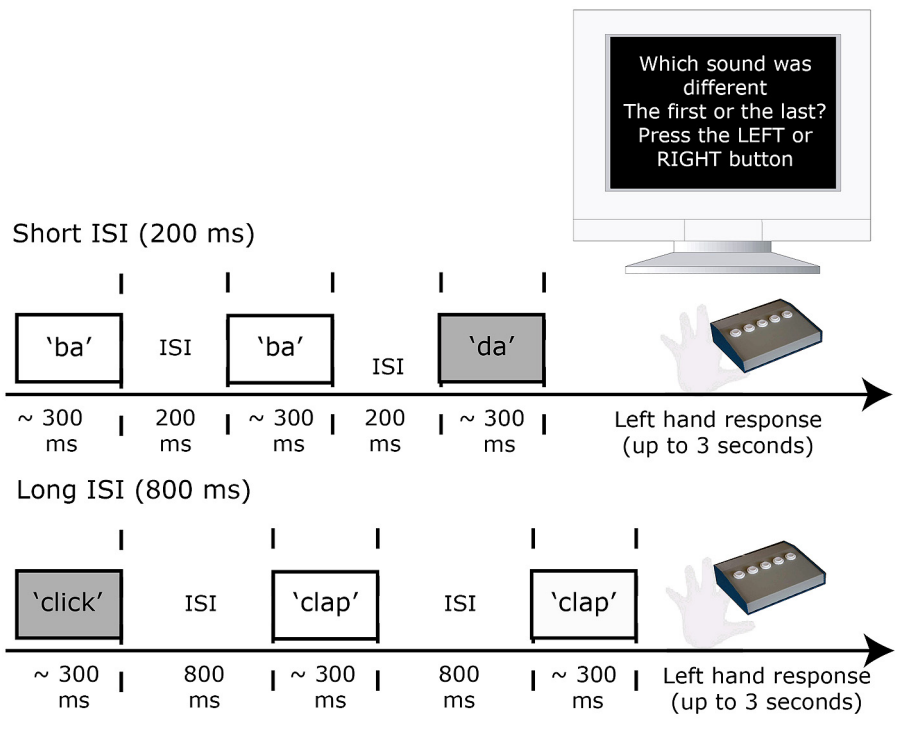


FIGURE 2 | Experimental design for the discrimination (AXB) task used pre and post stimulation. The across- and within-category sound triplets were heard at two ISIs (200 and 800 ms). Responses were made via a button-press response with the left hand. After 3 s without response the next trial began.

current flowing posterior–anterior under the junction of the two wings of the coil. The position and angle of the coil over the lateral surface was adjusted until a reliable MEP was observed in the targeted contralateral muscle (Möttönen et al., 2014). TMS was applied according to current safety guidelines (Wassermann, 1998; Rossi et al., 2009) with all participants required to complete a TMS safety screening form based on recommendations from Rossi et al. (2009). As there are no safety guidelines for the use of cTBS currently, the protocol of Huang et al. (2005) was followed.

Continuous theta-burst stimulation (cTBS)

Theta-burst TMS was applied continuously for 40 s over either the lip or hand representation of M1 cortex. The train of stimulation comprised 600 pulses, in high-frequency (50 Hz) triplets repeated at 5 Hz. The aMT for each participant was determined using the Magstim Super Rapid^2 stimulator as the intensity at which single TMS pulses elicited more than 5 out of 10 MEPs with amplitude of at least 200 μV when the muscle was contracted at 20–30% of the maximum. The aMT was determined whilst participants maintained voluntary contraction of the hand or lip muscle at 20–30% of the maximum for 5 min. This was based on previous findings revealing that continuous contraction of the target muscle for 5 min influences the after-effects of cTBS (e.g., Gentner et al., 2008; Iezzi et al., 2008). Using the MagStim Super Rapid^2, the average aMT (percentage of maximum stimulator output, ± SEM) for the lip area was 65.17% (±2.41%) and for the hand area was 51.55% (±2.15%). cTBS was delivered at an intensity of 80% aMT while participants relaxed their lip and hand muscles. This ensured that intensities used were sub-threshold and, therefore, not strong enough to elicit MEPs at rest. This was confirmed by administering 10 single-pulses of TMS at 80% aMT while the lip and hand muscles were relaxed. No MEPS were observed in participants at the intensity of stimulation used to apply cTBS. The maximum possible theta-burst intensity of 50% maximum stimulator output was applied if the intensity at 80% aMT was greater than 50%. The mean intensity used during cTBS over the lip area of M1 was 49.08% (±0.47%). The mean intensity used during cTBS over the hand area was 41.18% (±1.7%). Following cTBS, participants were told not to contract their lip or hand muscles until after the experiment had finished as activation during or following cTBS has previously been shown to alter the after-effects (Huang et al., 2008).

Single pulse TMS

To assess the suppressive effects of cTBS on cortical excitability, single-pulse TMS was used to elicit MEPs from the target (lip or hand) and non-target (lip or hand) muscle before cTBS and 5, 10, 20, 25, 35, and 45 min later for comparison with MEPs collected pre-cTBS. Twenty-four MEPs were acquired prior to cTBS and at each time point post-cTBS (12 MEPs per muscle) with an inter-pulse interval ranging from 5 to 6.5 s ($M = 5.75$, SD $= 0.65$). MEPs were acquired from the lip muscle first followed by the hand muscle in all cases. The intensity used to administer the single pulses of TMS in each participant was determined using a MagStim200 prior to the aMT described above. The intensity was defined as that which produced MEPs with average peak-to-peak amplitude of 0.3 mV or 1 mV on 10 consecutive trials for the lip and hand muscles, respectively (Möttönen and Watkins, 2009; Murakami et al., 2011; Möttönen et al., 2014). All MEPs before and after cTBS were recorded from the relaxed muscles. The average intensity (±SEM) used to elicit MEPs in the lip muscle was 64% (±1.59%) and 53% (±3.86%) in the hand muscle. We note that the stimulator outputs differ between the Magstim 200 that generates monophasic pulses and the Super Rapid2 that generates biphasic pulses. Therefore, the percentage (%) of stimulator output used to elicit MEPs in a relaxed muscle and for the aMT in a contracted muscle on each of these stimulators is not comparable.

STATISTICAL ANALYSIS

For the behavioral data from the discrimination task, anticipation responses that were shorter than 200 ms were removed from the data (0.35% of total responses). If the participant did not respond within three seconds, then the next trial began (1.3% of responses were missed). Percent correct AXB responses for the across- and within-category stimuli were calculated for each contrast and each ISI separately. The scores post-cTBS were averaged across three time-bins; an early time bin (8–20 min post cTBS), a middle time-bin (20–35 min post cTBS) and a late time bin [35 min post cTBS–completion of the experiment (between 45 and 50 min)]. Missing data were replaced with the group mean for that contrast and ISI to allow the full ANOVA to be carried out (missing data occurred only at the late post-cTBS time point; 5/144 responses for the lip group and 12/132 responses for the hand group). For the two groups of participants who received lip ($n = 12$) and hand ($n = 11$) stimulation, two separate repeated-measures ANOVAs were carried out for the across- and within-category data. Within-subject effects of time (four levels: pre-cTBS, early, middle and late post-cTBS), ISI (200 vs. 800 ms), and stimulus type (three types: lip- and tongue-articulated and non-speech continua) were evaluated. *Post hoc* pairwise comparisons were used to compare time-points for the separate continua and ISI and were corrected using Bonferroni correction.

For the MEP data, MEPs with peak-to-peak amplitude greater than two SDs from the mean at each separate time point were removed as outliers (2.8% of responses). The remaining MEPs were averaged for the pre-cTBS time point, and the early post-cTBS (5 and 10 min), middle post-cTBS (20 and 25 min) and late post-cTBS (35 and 45 min) time points. A repeated-measures ANOVA with time (four levels: pre-cTBS, early, middle and late post-cTBS) as a within-subject factor was used to evaluate the effects of cTBS on motor cortex excitability for the lip and the hand data separately.

RESULTS

CATEGORICAL PERCEPTION OF SPEECH AND NON-SPEECH SOUNDS

Categorical perception of audio-morphed speech and non-speech continua averaged across all participants tested in session 1 ($n = 23$) is shown in **Figure 3**. The category boundary position for all stimulus continua in all participants was between the 35% and 65% along the acoustic continuum. Analysis of the slopes (β_1)

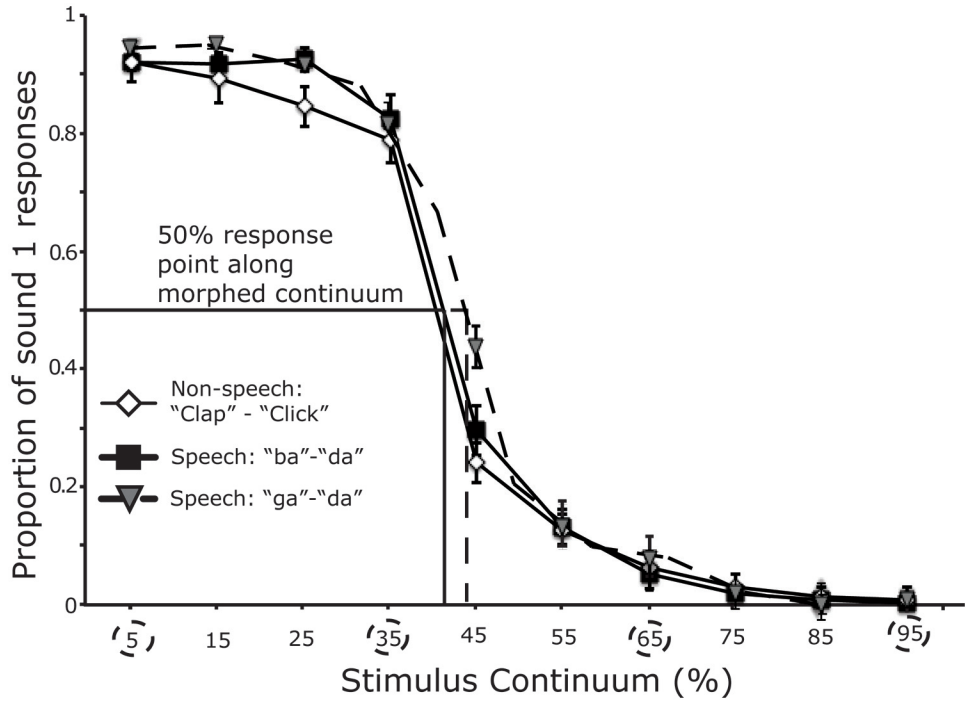


FIGURE 3 | Performance in the identification task (first session) across all participants. A logistic curve was fit to the raw data points across participants. The circles depict the points along the continuum defined as within- or across-category. Lines on the *x*-axis mark the position of the category boundary for each stimulus pair. Error bars represent the SE of the mean.

across stimulus continua revealed no significant difference in the steepness of the logistic curves ($F < 1$; "ba"–"da" = 0.90 ± 0.02; "ga"–"da" = 0.90 ± 0.02; "clap"–"click" = 0.92 ± 0.02). There was no significant difference in boundary position across stimulus pair ($F < 1$), ("ba"–"da": mean = 43.12 ± 1.18% da; "ga"–"da": mean = 44.39 ± 1.07% da; "clap"–"click": mean = 42.33 ± 1.01% click).

The data from the discrimination task that was performed before cTBS was combined for both groups of participants ($n = 23$) and analyzed using ANOVA with within-subjects factors of contrast (three types: ba–da, ga–da, and click–clap), stimulus type (across vs. within category) and ISI (200 vs. 800 ms); the between-subject factor of group was included but was not expected to be a main effect or interact significantly with any of the other factors. As expected for stimuli that are perceived categorically, accuracy on discrimination of across-category stimuli was significantly better than for within-category stimuli that had an equivalent acoustic difference (i.e., 30%) between them [$F(1,22) = 83.30$, $p < 0.0005$]. This main effect of stimulus type interacted significantly with the contrast, however [$F(2,42) = 4.83$, $p = 0.013$]; the main effect of contrast was significant also [$F(2,42) = 17.27$, $p < 0.0005$] due to significantly lower performance on the ga–da contrast compared with the other two contrasts (ba–da, $p = 0.002$, clap–click, $p < 0.0005$, corrected). The interaction between stimulus type and contrast was explored with separate ANOVAs for within and across category stimuli. This revealed a significant difference among the three contrasts for the scores on the within-category stimuli [$F(2,42) = 21.24$, $p < 0.0005$] but no difference among the scores for the across-category stimuli. The within-category stimuli were discriminated significantly more accurately for the non-speech clap–click contrast relative to the other two contrasts (ba–da: mean difference = 6.96 ± 2.00%, $p = 0.007$; ga–da: mean difference = 11.95 ± 1.85%, $p < 0.0005$); the within-category stimuli were also discriminated more accurately for the ba–da contrast relative to the ga–da contrast (mean difference = 5.00 ± 1.66%, $p = 0.020$; corrected; see **Figure 4**). The non-speech contrast was also the only one to show a significant difference in accuracy according to ISI [$F(1,22) = 5.58$, $p = 0.027$]; performance on the longer ISI was better than for the shorter ISI. The interaction between ISI and stimulus type was not significant, however.

In sum, all three types of contrast (ba–da, ga–da, and clap–click) were perceived categorically before cTBS was applied to either the hand or the lip representation in M1. Although there were no significant differences among the slopes of the identification functions for the three different contrasts, the better performance on the within-category discrimination of the clap–click contrast relative to the other two contrasts suggests that the non-speech contrast was perceived less categorically than the speech contrasts (see **Figure 4**).

THE EFFECT OF cTBS ON DISCRIMINATION OF SPEECH AND NON-SPEECH SOUNDS

Discrimination of across-category stimuli

Change in discrimination accuracy due to cTBS for across-category stimuli was evaluated using ANOVA with within-subject

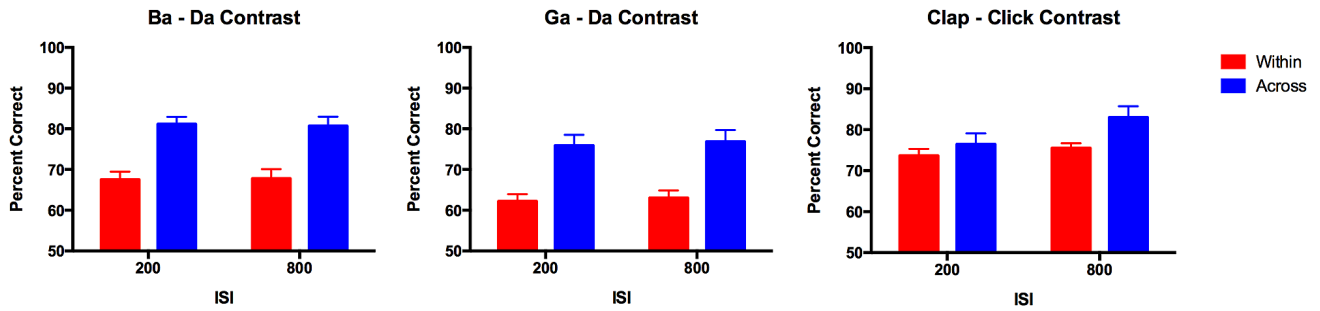


FIGURE 4 | Pre-stimulation discrimination accuracy for across and within-category stimuli. Bars represent the group mean ($n = 23$); error bars represent the SE of the mean. The three contrasts each showed the typical pattern associated with categorical perception wherein across category pairs are discriminated more robustly than within-category pairs. This pattern is weakest for the non-speech contrast but was still significant.

factors of contrast (three types: ba–da, ga–da, clap–click), time (four time-points: pre-cTBS, early, middle, and late post-cTBS), and ISI (200 vs. 800 ms). For the group of participants that received cTBS over the lip representation ($n = 12$), there was a close to significant three-way interaction following a Greenhouse–Geisser correction to the degrees of freedom because of non-sphericity of data (Mauchly's test of sphericity, $p = 0.036$; $F(3.46,38.03) = 2.258$, $p = 0.089$). The two-way interaction between contrast and time was significant [$F(6,66) = 2.355$, $p = 0.040$] as was the main effect of time [$F(3,33) = 6.144$, $p = 0.002$]. Separate ANOVAs for the three different contrasts revealed that the two-way interaction between contrast and time was due to a significant effect of time for the "ba"–"da" speech contrast [$F(3,33) = 9.291$, $p < 0.0005$] and not for the other two contrasts (ga–da and clap–click). *Post hoc* pairwise comparisons revealed the main effect of time in the "ba"–"da" contrast was due to a significant reduction in performance at the middle post-cTBS time point relative to all others (mean difference ± SEM for middle post-cTBS compared with: pre-cTBS = 15.54 ± 2.83%, $p = 0.001$; early post-cTBS = 18.29 ± 3.98%, $p = 0.005$; late post-cTBS = 15.34 ± 4.07%, $p = 0.019$; p-values corrected). This effect was greater for triplets presented with ISI of 800 ms than for those with ISI of 200 ms at the middle time point [$t(11) = 3.18$, $p = 0.009$; **Figure 5A**]. Note, however, that the time by ISI interaction for the "ba"–"da" contrast did not quite meet the $p < 0.05$ cutoff for significance [$F(3,33) = 2.748$, $p = 0.058$]. One-sample t-tests were also used to test if discrimination of "ba"–"da" stimuli was above chance (50%) at the middle time point. Only the discrimination of "ba"–"da" stimuli presented with a short ISI (200 ms) was significantly above chance [$t(11) = 5.90$, $p < 0.0005$]; discrimination of triplets presented with a longer ISI (800 ms) did not differ from chance performance ($p = 0.140$).

For the group of participants that received cTBS over the hand representation ($n = 11$), there were no significant main effects or interactions (**Figure 5B**); there was a close-to-significant interaction between contrast and ISI [$F(2,20) = 3.40$, $p = 0.054$], which appears to be due to better performance at the longer ISI for the non-speech ("clap"–"click") contrast across all time-points (see also results above for the pre-TBS data analysis).

Discrimination of within-category stimuli

Change in discrimination accuracy due to cTBS for within-category stimuli was evaluated using ANOVA as described above with within-subject factors of contrast (three types: ba–da, ga–da, clap–click), time (four time-points: pre-cTBS, early, middle, and late post-cTBS), and ISI (200 vs. 800 ms). For the group of participants that received cTBS over the lip representation ($n = 12$), there was a significant interaction between contrast and time [$F(6,66) = 2.37$, $p = 0.039$] and a significant main effect of contrast [$F(2,22) = 10.50$, $p = 0.001$]. The main effect of contrast was due to lower performance on the within-category discrimination for the "ga"–"da" contrast compared to the non-speech "clap"–"click" contrast (mean difference = 6.97 ± 1.37%, $p = 0.001$), which suggests a more typical categorical perception performance for the speech compared to the non-speech stimuli (see results above for the pre-TBS data analysis). Separate ANOVAs for each of the three different contrasts showed a significant main effect of time [$F(3,33) = 6.83$, $p = 0.001$] for the "ga"–"da" contrast but not for the other speech ("ba"–"da") nor the non-speech ("clap"–"click") contrasts. The discrimination of within-category stimuli was significantly better at the late post-cTBS time point compared to the pre-cTBS time point (mean difference = 9.29 ± 1.99%, $p = 0.004$, corrected) and the early post-cTBS time point (mean difference = 7.97 ± 2.38%, $p = 0.039$, corrected), indicating improved performance over the course of the experiment in discriminating within-category stimuli for the "ga"–"da" speech contrast (see **Figure 6A**). For the group of participants that received cTBS over the hand representation ($n = 11$), there was a significant main effect of contrast [$F(2,20) = 12.22$, $p < 0.0005$]; none of the other main effects or interactions were significant, though the main effect of ISI was close [$F(1,10) = 4.01$, $p = 0.073$; **Figure 6B**].

Summary of discrimination results

In sum, cTBS over the lip but not the hand representation in M1 significantly reduced the ability of participants to discriminate speech sounds that are lip articulated from those that are tongue articulated but not their ability to discriminate speech sounds from different phonetic categories that are both tongue articulated nor non-speech sounds made by the hands. The reduction

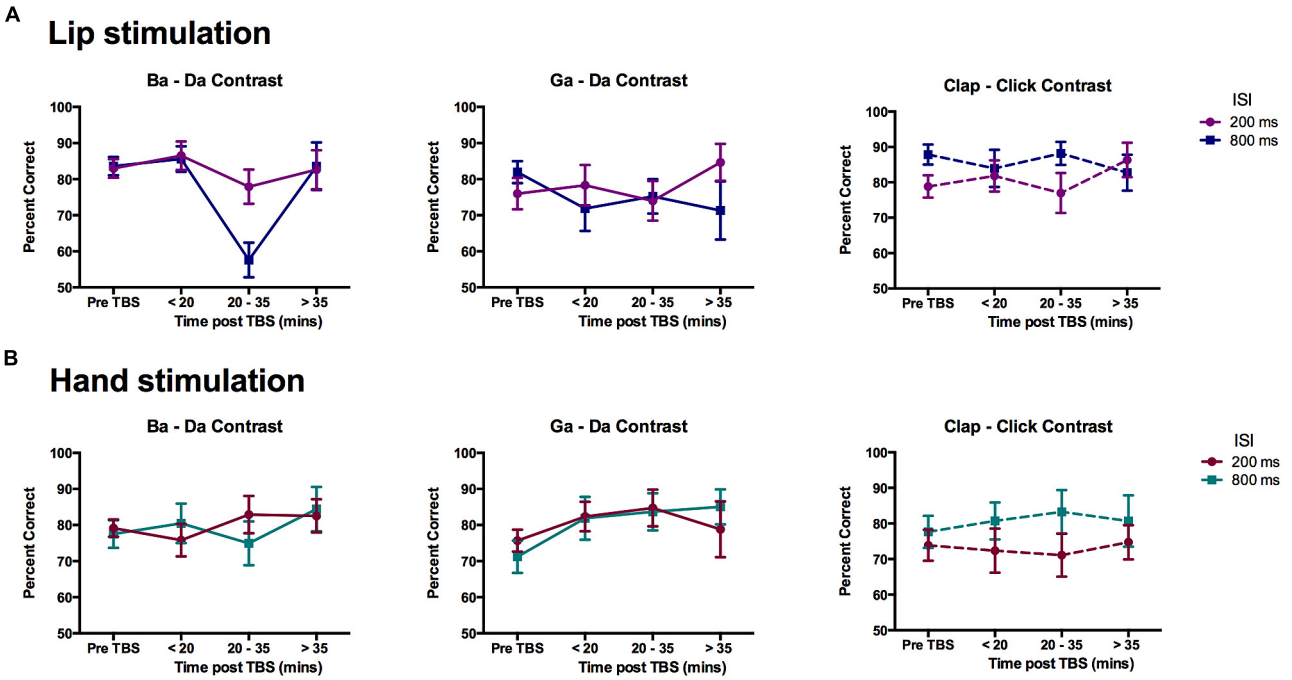


FIGURE 5 | Effects of cTBS on discrimination of across-category pairs. Mean percent correct scores for the participants in the **(A)** Lip stimulation group ($n = 12$) and **(B)** Hand stimulation group ($n = 11$). Data are plotted separately for each contrast. The graphs show the data for each time point in the experiment with the two ISI plotted as separate lines. Error bars represent the SE of the mean. The only significant reduction in performance was seen for the data obtained between 20 and 35 min post-cTBS to the lip representation for the lip-articulated "ba"–"da" contrast.

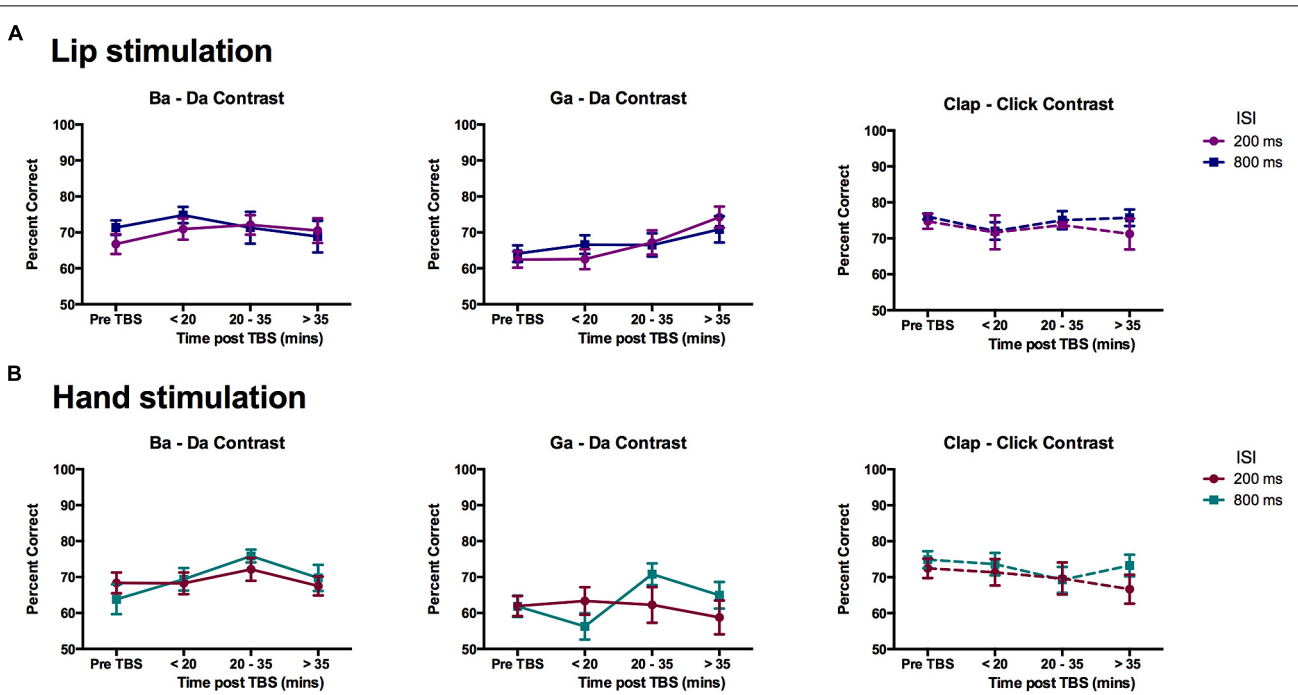


FIGURE 6 | Effects of cTBS on discrimination of within-category pairs. Mean percent correct scores for the participants in the **(A)** Lip stimulation group ($n = 12$) and **(B)** Hand stimulation group ($n = 11$). A significant improvement in discrimination accuracy for the "ga"–"da" contrast in the group who received cTBS to the lip representation can be seen in the middle graph of the top row. In the hand stimulation group, the performance on the "ga"–"da" contrast was significantly lower than that for the non-speech contrast.

in discrimination ability was observed only at a time point occurring 20 min after the stimulation and was not seen in the data obtained earlier at 5 and 10 min post stimulation or later at 35 and 45 min post stimulation. Discrimination of the lip-articulated speech sounds dropped to chance level at this middle time-point when they were presented with an ISI of 800 ms, whereas discrimination performance for stimuli presented with an ISI of 200 ms was slightly reduced but remained significantly above chance.

THE EFFECT OF cTBS ON MOTOR EXCITABILITY

For the data obtained from the lip target muscle, there was no significant effect of cTBS on MEP size [$F(3,33) = 1.34$, $p = 0.277$]. Similarly, cTBS over the hand representation, had no significant effect on MEP size recorded from the hand target muscle [$F(3,30) = 2.11$, $p = 0.120$; **Figure 7**]. Analysis of MEPs recorded from the non-target muscle also revealed no significant change in motor excitability, $F < 1$. In sum, for the group data, 40-s cTBS over either the lip or hand representation in M1 did not significantly change motor excitability in either area as indexed by the size of MEPs elicited by single pulse TMS. Nevertheless, we wished to explore whether the reduction in discrimination ability seen at the time point occurring 20 min after the stimulation was related to the efficacy of cTBS to reduce MEP size in some of the participants. Five participants showed a decrease in MEP size at the middle time point relative to the pre-cTBS MEP size (3.5–16% reduction) and seven participants did not. Performance of these two subgroups on the discrimination of the "ba"–"da" speech contrast at the two ISIs was compared using independent *t*-tests. There were no significant differences. However, the subgroup showing decreased motor excitability had a lower mean performance (49.8%; i.e., chance) on discrimination of the stimuli at the longer ISI compared to the subgroup that did not show a reduction in MEP size (63.2%).

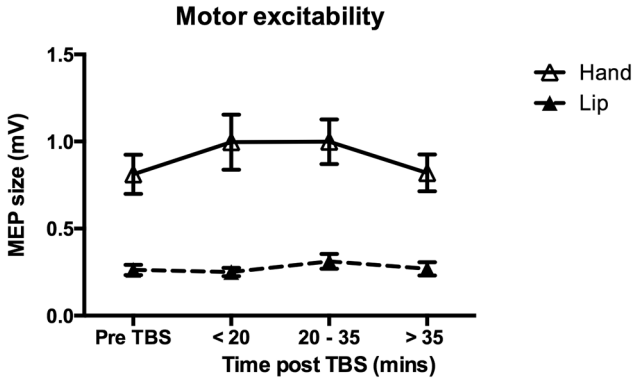


FIGURE 7 | Effect of cTBS on motor excitability. Group mean MEP sizes at each time point (pre-cTBS, early, middle and late post-cTBS) are shown for the target muscle in response to single-pulse TMS over either the Lip (dashed line) or the Hand (solid line) representation in M1. Error bars represent SEs of the mean. There was no significant change in excitability due to cTBS in either the Lip ($n = 12$) or the Hand ($n = 11$) group. The difference in amplitudes for the MEPs elicited in the Lip compared to the Hand muscle was expected as this was used in determining the thresholds separately for each muscle (see Section "Materials and Methods" for details).

DISCUSSION

To our knowledge, this is the first study to assess the effects of cTBS on discrimination of naturally recorded speech and non-speech sounds. By applying 40 s of cTBS over the lip representation in the motor cortex, we temporarily impaired the ability of listeners to discriminate syllables from different phonetic categories on a continuum that varied in place of articulation from the lips to the tongue ("ba" to "da"); the ability to discriminate syllables from within a phonetic category was unaffected by cTBS. The impairment observed was maximal between 20 and 35 min after the stimulation for the across-category "ba"–"da" stimuli presented with a longer ISI (800 ms) and, in fact, discrimination performance at this time was reduced to chance levels. Discrimination of the same stimuli presented with a short ISI (200 ms) at the same time-point was slightly less affected by the cTBS and remained above chance. The finding that the impairment occurred during a 15-min time period starting 20 min after cTBS was applied is consistent with the time period at which the maximum inhibitory effects of cTBS on motor excitability have been previously reported (Huang et al., 2005). Data obtained earlier (i.e., in the first 20 min following stimulation) and later (i.e., more than 35 min after the stimulation) did not show any changes in discrimination ability relative to the pre-stimulation baseline data. This impairment in discrimination of speech sounds was not seen when the hand representation of the motor cortex was stimulated. These results support previous studies that have shown a mediating role of the motor system in speech perception with performance in speech tasks significantly affected following TMS over the motor regions (e.g., Meister et al., 2007; D'Ausilio et al., 2009; Sato et al., 2009; Bartoli et al., 2013). These findings also replicate our previous results using low frequency (0.6 Hz) repetitive TMS for 15 min to temporarily disrupt function in the lip representation of primary motor cortex, impairing categorical perception and discrimination of speech syllables that involved the lips in their production (Möttönen and Watkins, 2009). There were a number of important differences between the two studies, however, and these are discussed below.

Firstly, 40-s of cTBS was used in the current study because we anticipated that this would induce a longer-lasting disruptive effect than that induced by 15 min of low frequency repetitive TMS, which we used previously. We found that the short train of high-frequency stimulation was well tolerated by participants and the behavioral disruption lasted for about 15 min occurring 20 min after the train had ended. The timing and duration of the behavioral effect requires replication but the technique in general offers some useful potential applications in future studies on the neural basis of speech processing. For example, TBS has been used successfully in combination with paired pulse TMS to study connectivity during speech processing (Murakami et al., 2012). It might also be combined usefully with neuroimaging techniques to further investigate auditory-motor processing of speech sounds (see Möttönen et al., 2013, 2014).

Another important difference from our previous study is that in the current study we used natural speech sounds recorded from three different speakers and audio-morphed into continua using the "Straight" channel-vocoder (Kawahara et al., 1999). Previously,

we used computer-generated artificial speech syllables and created a single continuum for each contrast by changing the slope of the formant transitions. Also, in the current experiment we included a novel non-speech contrast, creating three continua based on sounds made by the hands ("clapping") and fingers ("clicking"), which were also perceived categorically. Including these stimuli allowed us to test for a possible double dissociation, whereby stimulation over the hand area might disrupt categorical perception of sounds made with the hands and not speech sounds whilst stimulation over the lip area might disrupt categorical perception of sounds made with the lips and not those made with the hands. Unfortunately, our findings are consistent with a single dissociation only, namely that the lip stimulation affected perception of speech sounds that were lip articulated and had no effect on the perception of sounds made by the hands; the hand stimulation did not impair discrimination of any auditory stimuli.

In the current study, the behavioral testing implemented used an AXB-type discrimination design with all stimulus sounds presented in a random order mixing the contrasts and continua tested. These included two different speech continua each from three speakers and three non-speech continua. Using the AXB discrimination task addresses criticism of our previous findings using low frequency rTMS and a same-different paradigm (Möttönen and Watkins, 2009) relating to the possibility that response bias changed rather than speech perception (Hickok, 2010). The previous study did not include identical pairs in the same–different task, which meant we could not evaluate changes in response bias using signal detection theory. The impaired discrimination of speech sounds reported in the current study cannot be explained by a change in response bias lending further support to our original claim that stimulation of the motor cortex impairs speech perception.

A final difference between the two studies was that we tested two different ISIs in the current experiment (200 vs. 800 ms), whereas we used 500 ms between stimulus pairs in our previous same-different task. Discrimination accuracy of lip- ("ba") versus tongue-articulated ("da") speech syllables was reduced to chance level 20 min after the cTBS over the lip representation only when the sound stimuli were presented at the longer ISI of 800 ms. Note, however, that the three-way interaction between contrast, time, and ISI was not quite significant and that discrimination of the same stimuli presented at this timepoint with 200 ms ISI was slightly affected by cTBS also. This is a novel finding and is consistent with evidence suggesting that once auditory memory has faded listeners must rely on pre-established phonetic representations to distinguish between speech sounds (Pisoni, 1973; Massaro and Cohen, 1983; Gerrits and Schouten, 2004). We propose this difference reflects phonetic vs. acoustic perception; at the shorter ISI, participants are more reliant on auditory "echoic" memory whilst at the longer ISI the auditory information is lost and participants are reliant on pre-established phonetic categories. It is this ability that is impaired when discriminating between lip- ("ba") versus tongue-articulated ("da") speech sounds following cTBS over the lip motor representation. This is consistent with studies assessing the role of verbal working memory and articulatory rehearsal in phonological discrimination (Boatman, 2004; Gough et al., 2005; Romero et al., 2006; Sato et al., 2009). Compared to sham stimulation, rTMS applied over left ventral premotor cortex significantly disrupted the ability to perform phoneme discrimination (Romero et al., 2006) and phonemic segmentation (Sato et al., 2009). One interpretation is that rTMS temporarily disrupts the recruitment of articulatory-based motor representations during phonological processing that are dependent on phonemic segmentation cues and a phonological short-term working memory store (Baddeley, 1990; Zatorre et al., 1992; Burton et al., 2000).

We report no effect of cTBS over the lip motor representation on discrimination accuracy for sounds that do not require the lips for articulation ("da" vs. "ga") and for non-speech sounds ("clap" vs. "click"). This shows that the impairment was specific to the articulatory features of the speech sounds. cTBS over the hand motor representation also had no effect on discrimination accuracy of speech or non-speech sounds showing that the temporary inhibition was specific to the lip rather than the hand motor representation. These results support previous studies investigating the contribution of articulatory motor cortex to perceptual speech processing and are consistent with claims that the lip motor representation contributes to speech perception in an articulator-specific manner (e.g., Möttönen and Watkins, 2009).

We also investigated the effects of applying 40 s of cTBS over the lip and hand representation of M1 on motor excitability, with 40 s of continuous stimulation shown to be more robust in inducing an inhibitory effect than protocols using 20 s of cTBS (e.g., Gentner et al., 2008). We found no significant inhibitory or facilitatory effect of cTBS over the lip or hand motor representation on MEPs recorded from the lip or hand target muscle. Thus, we did not replicate findings from previous studies revealing an inhibitory effect of 40 s of cTBS on motor excitability (e.g., Huang et al., 2005, 2008; Gentner et al., 2008). One possible account for why no effect of cTBS on the size of MEPs was observed in our study is that we recorded MEPs alongside the discrimination responses. Whilst all behavioral responses were made with the left hand, ipsilateral to the site of stimulation to avoid motor excitability changes due to hand movements, interhemispheric inhibition from the right motor cortex cannot be ruled out as affecting the left motor cortex excitability in an unexpected way. Increased attentional demands present during discrimination of the speech and non-speech sounds may also have contributed to the absence of an effect of cTBS on motor excitability.

A more likely explanation of our failure to replicate previous findings of reduced motor excitability following cTBS relates to recent reports of highly variable responses to cTBS across protocols and across participants. For example, applying cTBS for 20 s over left M1 facilitated rather than suppressed the amplitude of MEPs recorded from the contralateral hand (Gentner et al., 2008). Suppressed motor excitability occurred only when voluntary muscle contraction was performed before cTBS. By doubling the duration of stimulation (applying cTBS for 80 s instead of 40 s). Gamboa et al. (2010) found a reversed facilitatory rather than inhibitory effect showing that the latter is not increased by simply prolonging the period of stimulation. Recently, Hamada et al.

(2013) also failed to replicate the suppression of motor excitability in a large group of healthy volunteers. They reported high inter-individual variability, which has been attributed to potential differences among individuals in the excitability of populations of neurons activated following cTBS (Day et al., 1987; Rothwell, 1997; Ridding and Ziemann, 2010; Hamada et al., 2013). A number of potential factors have been suggested that contribute to this variability including age, gender, time of day, hormonal influence (e.g., changes in cortisol levels), neuromodulators and genetics (Ridding and Ziemann, 2010). A systematic investigation of inter-individual variability for theta-burst protocols reported no consistent pattern of response among individuals related to age, gender, time of day or initial differences in stimulation intensity thresholds and baseline MEP amplitude (Hamada et al., 2013). Rather, Hamada et al. (2013) suggested that the inter-individual variation observed reflects differences between people in the population of neurons activated by theta-burst stimulation that might be determined by differences in cortical anatomy.

In the current study, we examined whether individuals who showed a reduction in motor excitability (as indexed by MEP amplitude changes) also showed a greater behavioral impairment. There was a trend in the data to support this view but the two subgroups of "responders" ($n = 5$) and "non-responders" ($n = 7$) were not significantly different in their ability to discriminate stimuli at the middle post-cTBS time-point when as a group they showed a significant decrement in task performance. Taking into account our own experience and the confusion in the literature, it is possible that MEPs are not always reliable indicators of the efficacy of cTBS on motor excitability.

CONCLUSION

Using cTBS, we replicated our previous findings that temporary disruption of the lip motor representation impairs the perception of speech sounds that rely on the lips for their production. This impairment is not explained by a change in response bias as it was obtained using an AXB discrimination task. Furthermore, we found that the effect of the TMS-induced disruption occurs predominantly for discrimination that relies on pre-existing phonetic categories and affects discrimination that relies on shorter-term acoustic representations to a lesser extent. This novel finding arose from a longer behavioral testing session with a larger number of natural speech and non-speech continua that was afforded by the anticipated longer-lasting effects of cTBS relative to low-frequency rTMS. A further advantage of TBS is that this longer-lasting effect is brought about by a very brief stimulation train (40 s compared to 15 min of low frequency rTMS). The use of TBS for further studies of speech processing holds promise, therefore. The effect of cTBS on motor excitability in our study was negligible, however. Although this failure to replicate previous effects was unexpected, the literature supports a picture of high inter-individual variability in motor excitability changes in response to TBS. It is as yet unknown whether similar variability affects behavioral responses. Our findings suggest, however, that cTBS over the motor cortex can affect behavior even when changes in motor excitability are not reliable.

ACKNOWLEDGMENTS

This study was funded by the Wellcome Trust WT091070AIA. Jack C. Rogers was supported by a Wellcome Trust Project Grant awarded to Kate E. Watkins and Riikka Möttönen. Riikka Möttönen was supported by Medical Research Council, U.K.

REFERENCES

Baddeley, A. D. (1990). *Human Memory: Theory and Practice.* London: Lawrence Erlbaum Associates.

Bartoli, E., D'Ausilio, A., Berry, J., Badino, L., Bever, T., and Fadiga, L. (2013). Listener–speaker perceived distance predicts the degree of motor contribution to speech perception. *Cereb. Cortex* 24, 1–8.

Boatman, D. F. (2004). Cortical bases of speech perception: evidence from functional lesion studies. *Cognition* 92, 47–65. doi: 10.1016/j.cognition.2003.09.010

Burton, M. W., Small, S. L., and Blumstein, S. E. (2000). The role of segmentation in phonological processing: an fMRI investigation. *J. Cogn. Neurosci.* 12, 679–690. doi: 10.1162/089892900562309

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The Motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

Day, B. L., Rothwell, J. C., Thompson, P. D., Dick, J. P., Cowan, J. M., Berardelli, A., et al. (1987). Motor cortex stimulation in intact man. 2. Multiple descending volleys. *Brain* 110(Pt 5), 1191–1209. doi: 10.1093/brain/110.5.1191

Devlin, J. T., and Watkins, K. E. (2007). Stimulating language: insights from TMS. *Brain* 130, 610–622. doi: 10.1093/brain/awl331

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402. doi: 10.1046/j.0953-816x.2001.01874.x

Gamboa, O. L., Antal, A., Moliadze, V., and Paulus, W. (2010). Simply longer is not better: reversal of theta burst after-effect with prolonged stimulation. *Exp. Brain Res.* 204, 181–187. doi: 10.1007/s00221-010-2293-4

Gentner R., Wankerl K., Reinsberger C., Zeller D., and Classen, J. (2008). Depression of human corticospinal excitability induced by magnetic theta-burst stimulation: evidence of rapid polarity-reversing metaplasticity. *Cereb. Cortex* 18, 2046–2053. doi: 10.1093/cercor/bhm239

Gerrits, E., and Schouten, M. E. (2004). Categorical perception depends on the discrimination task. *Percept. Psychophys.* 66, 363–376. doi: 10.3758/BF03194885

Gough, P. M., Nobre, A. C., and Devlin, J. T. (2005). Dissociating linguistic processes in the left inferior frontal cortex with transcranial magnetic stimulation. *J. Neurosci.* 25, 8010–8016. doi: 10.1523/JNEUROSCI.2307-05.2005

Grossheinrich, N., Rau, A., Pogarell, O., Hennig-Fast, K., Reinl, M., Karch, S., et al. (2009). Theta burst stimulation of the prefrontal cortex: safety and impact on cognition, mood, and resting electroencephalogram. *Biol. Psychiatry* 65, 778–784. doi: 10.1016/j.biopsych.2008.10.029

Hamada, M., Murase, N., Hasan, A., Balaratnam, M., and Rothwell, J. C. (2013). The role of interneuron networks in driving human motor cortical plasticity. *Cereb. Cortex* 23, 1593–1605. doi: 10.1093/cercor/bhs147

Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. *Lang. Cogn. Process.* 25, 749–776. doi: 10.1080/01690961003595572

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Huang, Y. Z., Edwards, M. J., Rounis, E., Bhatia, K. P., and Rothwell, J. C. (2005). Theta burst stimulation of the human motor cortex. *Neuron* 45, 201–206 doi: 10.1016/j.neuron.2004.12.033

Huang, Y. Z., Rothwell, J. C., Edwards, M. J., and Chen, R. S. (2008). Effect of physiological activity on an NMDA-dependent form of cortical plasticity in human. *Cereb. Cortex* 18, 563–570. doi: 10.1093/cercor/bhm087

Iezzi, E., Conte, A., Suppa, A., Agostino, R., Dinapoli, L., Scontrini, A., et al. (2008). Phasic voluntary movements reverse the aftereffects of subsequent theta-burst stimulation in humans. *J. Neurophysiol.* 100, 2070–2076. doi: 10.1152/jn.90521.2008

Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Commun.* 27, 187–207. doi: 10.1016/S0167-6393(98)00085-5

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Liberman, A. M., Harris, K. S., Hoffman, H., and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide.* Cambridge: Cambridge University Press.

Massaro D. W., and Cohen, M. M. (1983). Categorical or continuous speech perception: a new test. *Speech Commun.* 2, 15–35. doi: 10.1016/0167-6393(83)90061-4

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Möttönen, R., Dutton, R., and Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cereb. Cortex* 23, 1190–1197. doi: 10.1093/cercor/bhs110

Möttönen, R., Rogers, J., and Watkins, K. E. (2014). Stimulating the lip motor cortex with transcranial magnetic stimulation. *J. Vis. Exp.* e51665. doi: 10.3791/51665

Möttönen, R., and Watkins, K. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825. doi: 10.1523/JNEUROSCI.6018-08.2009

Murakami, T., Restle, J., and Ziemann, U. (2011). Observation–execution matching and action inhibition in human primary motor cortex during viewing of speech-related lip movements or listening to speech. *Neuropsychologia* 49, 2045–2054. doi: 10.1016/j.neuropsychologia.2011.03.034

Murakami, T., Restle, J., and Ziemann, U. (2012). Effective connectivity hierarchically links temporoparietal and frontal areas of the auditory dorsal stream with the motor cortex lip area during speech perception. *Brain Lang.* 122, 135–141. doi: 10.1016/j.bandl.2011.09.005

Oberman, L. M., Edwards, D., Eldaief, M., and Pascual-Leone, A. (2011). Safety of theta burst transcranial magnetic stimulation: a systematic review of the literature. *J. Clin. Neurophysiol.* 28, 67–74. doi: 10.1097/WNP.0b013e318205135f

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Percept. Psychophys.* 13, 253–260. doi: 10.3758/BF03214136

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *J. Acoust. Soc. Am.* 61, 1352–1361. doi: 10.1121/1.381409

Pisoni, D. B., and Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *J. Acoust. Soc. Am.* 55, 328–333. doi: 10.1121/1.1914506

Pisoni, D. B., and Tash, J. (1974). Reaction-times to comparisons within and across phonetic categories. *Percept. Psychophys.* 15, 285–290. doi: 10.3758/BF03213946

Pulvermüller, F., and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* 11, 351–360. doi: 10.1038/nrn2811

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Repp, B. H. (1984). "Categorical perception: issues, methods, findings." in *Speech and Language: Advances in Basic Research and Practice*, Vol. 10, ed. N. J. Lass (Orlando, FL: Academic Press), 244–322.

Ridding, M. C., and Ziemann, U. (2010). Determinants of the induction of cortical plasticity by non-invasive brain stimulation in healthy subjects. *J. Physiol.* 588, 2291–2304. doi: 10.1113/jphysiol.2010.190314

Romero, L., Walsh, V., and Papagno, C. (2006). The neural correlates of phonological short-term memory: a repetitive transcranial magnetic stimulation study. *J. Cogn. Neurosci.* 18, 1147–1155. doi: 10.1162/jocn.2006.18.7.1147

Rossi, S., Hallett, M., Rossini, P. M., and Pascual-Leone, A. (2009). The safety of TMS consensus group. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin. Neurophysiol.* 120, 2008–2039. doi: 10.1016/j.clinph.2009.08.016

Rothwell, J. C. (1997). Techniques and mechanisms of action of transcranial stimulation of the human motor cortex. *J. Neurosci. Methods* 74, 113–122. doi: 10.1016/S0165-0270(97)02242-5

Roy, A. C., Craighero, L., Fabbri-Destro, M., and Fadiga, L. (2008). Phonological and lexical motor facilitation during speech listening: a transcranial magnetic stimulation study. *J. Physiol. Paris* 102, 101–105. doi: 10.1016/j.jphysparis.2008.03.006

Sato, M., Temblay, P., and Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7. doi: 10.1016/j.bandl.2009.03.002

Scott, S. K., McGettigan, C., and Eisner, F. (2009). A little more conversation, a little less action: candidate roles for motor cortex in speech perception. *Nat. Rev. Neurosci.* 10, 295–302. doi: 10.1038/nrn2603

van Casteren, M., and Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behav. Res. Methods* 38, 584–589. doi: 10.3758/BF03193889

Wassermann, E. M. (1998). Risk and safety of repetitive transcranial magnetic stimulation: report and suggested guidelines from the International Workshop on the Safety of Repetitive Transcranial Magnetic Stimulation, June 5–7, 1996. *Electroencephalogr. Clin. Neurophysiol.* 108, 1–16. doi: 10.1016/S0168-5597(97)00096-8

Watkins, K. E., Strafella, A. P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41, 989–994. doi: 10.1016/S0028-3932(02)00316-0

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Zatorre, R., Evans, A., Meyer, E., and Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science* 256, 846–849. doi: 10.1126/science.1589767

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2014; accepted: 27 June 2014; published online: 15 July 2014.
Citation: Rogers JC, Möttönen R, Boyles R and Watkins KE (2014) Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex. Front. Psychol. ***5****:754. doi: 10.3389/fpsyg.2014.00754*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 24 June 2014
doi: 10.3389/fpsyg.2014.00568

The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing

Lucie Scarbel*, Denis Beautemps, Jean-Luc Schwartz and Marc Sato

CNRS, Grenoble Images Parole Signal Automatique-Lab, Speech and Cognition Department, UMR 5216, Grenoble University, Grenoble, France

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Maurizio Gentilucci, University of Parma, Italy
Tim Saltuklaroglu, University of Tennessee Health Science Center, USA

****Correspondence:***
Lucie Scarbel, CNRS, Grenoble Images Parole Signal Automatique-Lab, Speech and Cognition Department, UMR 5216, Grenoble University, 1180 Avenue Centrale, BP 25, 38040 Grenoble Cedex 9, France
e-mail: lucie.scarbel@gipsa-lab.grenoble-inp.fr

One classical argument in favor of a functional role of the motor system in speech perception comes from the close-shadowing task in which a subject has to identify and to repeat as quickly as possible an auditory speech stimulus. The fact that close-shadowing can occur very rapidly and much faster than manual identification of the speech target is taken to suggest that perceptually induced speech representations are already shaped in a motor-compatible format. Another argument is provided by audiovisual interactions often interpreted as referring to a multisensory-motor framework. In this study, we attempted to combine these two paradigms by testing whether the visual modality could speed motor response in a close-shadowing task. To this aim, both oral and manual responses were evaluated during the perception of auditory and audiovisual speech stimuli, clear or embedded in white noise. Overall, oral responses were faster than manual ones, but it also appeared that they were less accurate in noise, which suggests that motor representations evoked by the speech input could be rough at a first processing stage. In the presence of acoustic noise, the audiovisual modality led to both faster and more accurate responses than the auditory modality. No interaction was however, observed between modality and response. Altogether, these results are interpreted within a two-stage sensory-motor framework, in which the auditory and visual streams are integrated together and with internally generated motor representations before a final decision may be available.

Keywords: speech perception, speech production, audiovisual speech perception, close-shadowing, sensorimotor interactions

INTRODUCTION

An old and classical debate in speech communication concerns the possible motor implication in speech perception and, more generally, the auditory vs. motor nature of the speech code. The heart of the debate relies in the existence and possible functional link between auditory and motor representations in both speech perception and speech production. Auditory theories of speech perception, such as the "Acoustic Invariance Theory" from Stevens and Blumstein (1978) or the "Adaptative Variability Theory" from Lindblom and Maddieson (1988) and Lindblom (1990) assume that speech perceptual processing and categorization are based on acoustic cues and auditory representations, with no need to call for any knowledge about the way the articulatory system produces the sound (Diehl et al., 2004). Conversely, the motor theory of speech perception (Liberman and Mattingly, 1985) and its direct realist variant (Fowler, 1986) claim that there is a crucial role of the motor system in speech perception, and consider that speech perception involves recovery of the stimulus cause, either physically (recovering the configuration of the vocal tract, in Fowler's direct realist theory) or biologically/cognitively (inferring motor commands in Liberman and Mattingly, 1985). More recently, a number of perceptuo-motor theories attempted various kinds of syntheses of arguments by tenants of both auditory and motor theories, proposing that implicit motor knowledge and motor representations are used in relationship with auditory representations and processes to elaborate phonetic decisions (Skipper et al., 2007; Schwartz et al., 2012).

It is worth noting that the question of whether articulatory processes mediate speech perception under normal listening conditions still remains vigorously debated (e.g., Hickok and Poeppel, 2007; Lotto et al., 2009; Scott et al., 2009; D'Ausilio et al., 2012; Schwartz et al., 2012). On the one hand, damage to motor speech areas in Broca's aphasic patients does not produce clear deficits in speech perception (e.g., Hickok et al., 2011) and studies using transcranial magnetic stimulation (TMS) also challenge a possible mediating role of the motor system in speech processing under normal listening conditions (Sato et al., 2009; D'Ausilio et al., 2011). On the other hand, an increasing number of neuroanatomical and neurophysiological studies suggest that there is indeed an active relationship between auditory and motor areas, both in speech perception and speech production. Indeed, brain imaging studies [functional magnetic resonance imaging (fMRI) or magnetoencephalography (MEG)] repeatedly showed the involvement of areas typically engaged in the speech production process (the left inferior frontal gyrus, ventral premotor cortex, primary motor cortex, somatosensory cortex) during various speech perception tasks (e.g., Binder et al., 2004; Möttönen et al., 2004; Wilson and Iacoboni, 2006; Grabski et al., 2013a), particularly in adverse conditions (e.g., noise: Zekveld et al., 2006; or foreign accent: Callan et al., 2004). TMS experiments confirmed the involvement of the motor system in speech perception, both auditory and audiovisual

(Fadiga et al., 2002; Wilson et al., 2004). However, evidence for a perceptuo-motor link in the human brain is not a proof that this link plays a functional role for processing speech inputs. Some neurophysiological evidence based on the use of TMS provided some evidence that perturbations of the motor system could lead to slight but significant modifications of the speech perceptual decision process (e.g., Meister et al., 2007; D'Ausilio et al., 2009; Möttönen and Watkins, 2009, 2012; Sato et al., 2009; Grabski et al., 2013b), but the perturbations are small and sometimes difficult to interpret.

In an influential review about the motor theory of speech perception, Galantucci et al. (2006) summarize different arguments to argue that "perceiving speech is perceiving gestures." One first argument comes from co-articulation effects and the fact that the acoustic properties of speech sounds are not invariant but context dependent. Since the correspondence between sounds and phonemes can be far from transparent, this led researchers to propose intended gestures as less invariant and as the ultimate objects of speech perception (see Liberman and Mattingly, 1985). Other arguments derive from close-shadowing effects and multisensory speech perception. Let us focus on these last two arguments, which will provide the basis for the present study.

Close-shadowing, which is an experimental technique in which subjects have to repeat speech immediately after hearing it, provides a natural paradigm for displaying perceptuo-motor links. In their pioneer study, Porter and Castellanos (1980) compared reaction times (RTs) in two speech perception tasks involving vowel-consonant-vowel (VCV) syllables (/aba/, /apa/, /ama/, /aka/, /aga/): in the first task, participants had to shadow the VCV they heard, that is to reproduce it orally as quickly as possible. They first produced the initial vowel and then shifted to the consonant as soon as they could perceive and identify it. The second task was a simple choice task: subjects had to shadow the initial vowel and, when stimulus changed into any consonant, they had to shift to /ba/ whatever the consonant, as quickly as possible. The authors found that RTs were of course faster in the simple task than in the shadowing task involving decision, but this difference was not very large (between 30 and 60 ms). Galantucci et al. (2006) compared those results with RTs found in Luce (1986), who used the same kind of paradigms (simple choice vs. multiple choice task, with comparable stimuli), but in responding by pressing a key rather than orally producing a response (in the choice task, participants had to press the key corresponding to the syllable they heard, and in the simple choice task they had to press a given key, whatever they heard). In Luce (1986) differences between RTs in the two tasks were larger than those in Porter and Castellanos's (1980) close-shadowing tasks (100/150 vs. 30/60 ms). This difference was interpreted by Galantucci et al. (2006) the assumption that, since perceiving speech is perceiving gestures, gesture perception will directly control speech response and make it faster. Later on, Fowler et al. (2003) published a study based on Porter and Castellanos's (1980) work, in which the participants had to shadow syllables in a "one choice task" and in a "multiple choice task" with three types of stimuli: /apa/, /aka/, and /ata/. In the "one choice task," participants were assigned to one of the three VCVs, shadowing the initial /a/ and instructed to switch toward their own consonant as soon as the stimulus consonant was presented, but independently of the identity of the stimulus consonant. In the "multiple choice task," participants simply had to shadow all VCVs. As in Porter and Castellanos (1980), they found that participants had shorter RTs in the simple choice task than in the multiple choice task. In the simple choice task, they also compared RTs between the three groups of subjects (one per assigned syllable) and they found that participants had shorter RTs when presented stimuli matched with their own syllable. These results are interpreted by Fowler et al. (2003), as well as by Galantucci et al., 2006, as suggesting that acoustic stimuli perceived as articulatory gestures would provide a prior "response goal" therefore modulating response times depending on the compatibility between stimulus and requested response.

Concerning multisensoriality on speech perception, it is known since long that lip-reading is helpful for understanding speech. Apart from the importance of lip-reading for hearing impaired subjects, normal-hearing subjects are able to lip-read (Cotton, 1935) and we know at least since Sumby and Pollack (1954) that the visual modality enhances auditory speech comprehension in noise. Shadowing experiments have actually also been exploited to assess audiovisual interactions in speech perception, though with no temporal constraint. Indeed, Reisberg et al. (1987) studied the audiovisual benefits in shadowing foreign language stimuli or linguistically complex utterances. In two experiments, he tested two groups of English participants to measure accuracy in production; participants were supposed to shadow French or German sentences, in audio vs. audiovisual conditions. Participants obtained significantly better scores – in terms of global accuracy of repetition – in the audiovisual condition compared with the audio condition. Then he tested one group of English participants who had to shadow English stimuli spoken with a Belgian accent, in audio and audiovisual conditions, in three experiments: one with simple phrases, one with more complex phrases and one with rare words. Once again, participants had better scores in the audiovisual condition. Then, Davis and Kim (2001) tested accuracy scores in repetitions of Korean phrases, by naïve English speakers, in a delayed shadowing experiment. Participants had to repeat stimuli at the end of the signal, in an audio and an audiovisual condition. After the repetition task, participants listened to a number of stimuli and had to decide whether they had already heard the stimuli or not. In both tasks, accuracy was better in the audiovisual condition.

However, all the audiovisual shadowing experiments do not deal with close-shadowing, hence they lack information about the dynamics of the decision process in relation with perceptuo-motor relationships. On the other side, close-shadowing experiments never involve audiovisual inputs, hence they lack information about the relationship between audiovisual interaction processes and perceptuo-motor interaction processes in phonetic categorization. Therefore audiovisual close-shadowing is the purpose of the present study in order to test audiovisual and perceptuo-motor interactions in an integrated paradigm.

One experiment was performed by two groups of French participants and focused on a comparative assessment of the accuracy and speed of oral vs. manual responses to auditory vs. audiovisual speech stimuli (VCV syllables). The speech stimuli were

presented without acoustic noise for the first group (Group A in the remainder of this paper) or with acoustic noise in the second one (Group B in the remainder of this paper). Our hypotheses were that (1) oral responses should be faster than manual responses, in agreement with previous studies on close-shadowing reported here above, and that (2) responses to audiovisual stimuli should be faster and more accurate than those to audio-only stimuli, at least in noise. An additional question concerns the possibility of interaction between these two components, evaluating whether the effect of vision is different from one modality of response (oral) to the other (manual). The responses to these questions will then be discussed in relationship with the debates about multisensory and perceptuo-motor interactions in speech perception.

MATERIALS AND METHODS

PARTICIPANTS

Two groups of respectively 15 and 14 healthy adults, native French speakers, participated in the experiment (Group A: 10 females; mean age: 29 years, age range: 20–39 years – Group B: 11 females; mean age: 24 years, age range: 19–34 years). All participants had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. The experiment was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

STIMULI

Multiple utterances of /apa/, /ata/, and /aka/ VCV syllables were individually produced by a male native French speaker (who did not participate in the experiment) in a sound-attenuated room. These three syllables were selected according to the distinct place of articulation of the consonant (stop bilabial /p/, alveolar /t/, and velar /k/) and to ensure a gradient of visual recognition between these syllables (with notably the bilabial /p/ consonant known to be more visually salient than alveolar /t/ and velar /k/ consonants). The syllables were audiovisually recorded using an AKG 1000S microphone and a high-quality digital video camera placed in front of the speaker zooming his face.

The corpus was recorded with the objective to obtain four different occurrences of /apa/, /ata/, and /aka/ with various durations of the initial /a/ vowel (i.e., 0.5s, 1s, 1.5s, and 2s). This was done in order to present participants with stimuli in which the onset of the consonant to categorize would occur at an unpredictable temporal position. To this aim, the speaker was asked to maintain the production of the initial vowel while expecting a visual "go" signal. The speaker produced 48 stimuli (4 initial durations × 3 types of syllables × 4 repetitions). One utterance was selected for each stimulus type and each initial vowel duration so as to obtain 12 stimuli. Then, to remove potential irrelevant acoustic differences between the stimuli, the occurrences of /apa/, /ata/, and /aka/ for a given expected initial duration were cut at their onset to equalize duration of the first vowel. Similarly, duration of the final vowel was equalized at 240 ms for all the 12 stimuli.

The audio tracks of the stimuli were sampled at 44.1 kHz and presented without noise in Group A. In Group B, the 12 stimuli were mixed with white noise, low pass filtered at −6 dB/oct, with a signal to noise ratio at −3 dB (the signal energy being defined from burst onset to the end of the vowel). In the audiovisual modality of the experiment, the video stream consisted in 572-by-520 pixel/images presented at a 50 Hz rate with the speaker's full face presented with blue lips to enhance lips movement perception.

EXPERIMENTAL PROCEDURE

The experiment consisted of two categorization tasks: close-shadowing in one case, where the responses were provided orally, by repeating as quickly as possible the presented speech syllables; manual decision in the other case, where the responses were provided manually, by pressing as quickly as possible the appropriate key. The stimuli to categorize consisted in /apa/, /ata/, and /aka/ syllables.

Participants were told that they would be presented with /apa/, /ata/, or /aka/ syllables, displayed either auditorily or audiovisually. In the close-shadowing task they were instructed to categorize and repeat each syllable as quickly as possible. To do so, they were asked to shadow the initial /a/ vowel and, when the stimulus changed to consonant, to immediately categorize and repeat the perceived CV syllable (/pa/, /ta/, or /ka/; see **Figure 1**). In the manual decision task, participants were instructed to categorize each syllable by pressing as quickly as possible with their dominant hand one of three keys respectively corresponding to /apa/, /ata/, or /aka/. The order of keys was counterbalanced across participants.

For each task (oral vs. manual response) and each modality (auditory vs. audiovisual), 16 occurrences of /apa/, /ata/, and /aka/syllables were presented in a fully randomized sequence of 48 trials. The order of task and modality of presentation was fully counterbalanced across participants.

Both groups performed the experiment in a soundproof room. Participants sat in front of a computer monitor at a distance of approximately 50 cm. The acoustic stimuli were presented at a comfortable sound level, with the same sound level set for all participants. While in Group A, the presentation of acoustic stimuli was done with a loudspeaker, the presentation of acoustic stimuli was done with earphones in Group B. This was required because of noisy stimuli, making acoustic processing complex and inaccurate if stimulus and response were mixed. The Presentation software (Neurobehavioral Systems, Albany, CA, USA) was used to control the stimulus presentation and to record key responses in the manual task. All participants' productions were recorded using an AKG 1000S microphone for off-line analyses, with a system ensuring synchrony between the stimulus presented to the participant and the participant's response. A brief training session preceded each task. The total duration of the experiment was around 30 min.

ACOUSTIC ANALYSES

In order to calculate RTs and the percentage of correct responses in the speech shadowing task, acoustic analyses of participants' productions were performed using Praat software (Boersma and Weenink, 2013). A semi-automatic procedure was first devised for segmenting participants' recorded productions. Based on minimal duration and low intensity energy parameters, the procedure involved the automatic segmentation of each utterance based on an intensity and duration algorithm detection. Then, for each presented stimulus, whatever the modality of presentation and

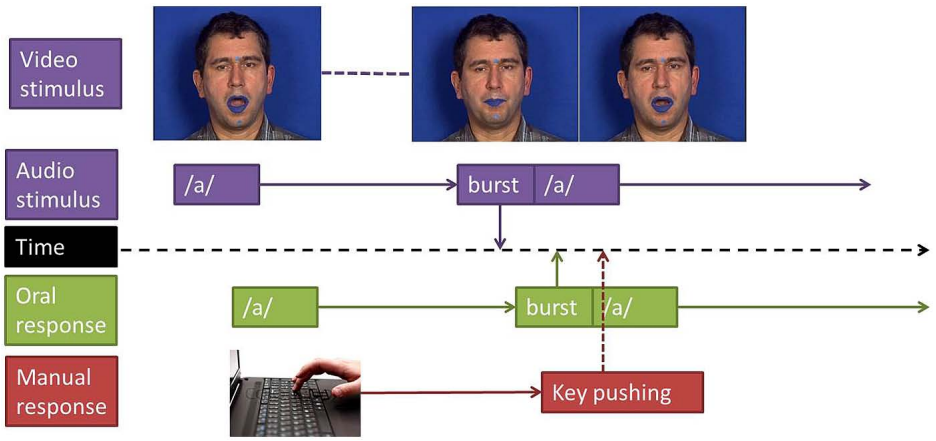


FIGURE 1 | Experimental design. Reaction times where measured between stimulus' and responses' burst (plosion release) for oral responses and between stimulus' burst and key pushing for manual responses.

response, an experimenter coded the participant's response and assessed whether it was correct or not.

Reaction times were estimated in reference to the burst onset of the stop consonant to categorize. In the manual decision task, the response instant was provided by the Presentation software, giving the instant when the key was pressed. In the close-shadowing tasks, the response time was provided by the burst onset of the stop consonant uttered by the participant in response to the stimulus, burst detection being realized by looking at the subject's production and inspecting waveform and spectrogram information with the Praat software. RTs were computed only for correct responses: omissions or any types of errors (replacing a consonant by another or producing two consonants or two syllables in the close-shadowing task) were excluded. The timelines of stimuli and responses, including description of the way response times were measured in both tasks, are displayed in **Figure 1**.

DATA ANALYSES

For each group, the percentage of correct responses and median RTs were individually determined for each participant, each task, each modality, and each syllable. Two repeated-measure ANOVAs were performed on these measures with the group (Group A with clear stimuli vs. Group B with noisy stimuli) as a between-subject variable and the task (close-shadowing vs. manual decision), the modality (auditory vs. audiovisual AV) and the syllable (/apa/ vs. /ata/ vs. /aka/) as within-subjects variables.

RESULTS

For all the following analyses, the significance level was set at $p = 0.05$ and Greenhouse–Geisser corrected (in case of violation of the sphericity assumption) when appropriate. All reported comparisons refer to *post hoc* analyses conducted with Bonferroni tests.

REACTION TIMES

As expected, the main effect of group was significant [$F(1,27) = 24.38$; $p < 0.001$], with faster RTs observed for clear stimuli in Group A compared to noisy/ambiguous stimuli in Group B (351 vs. 484 ms). Crucially, the main effects of task [$F(1,27) = 151.70$; $p < 0.001$] and modality [$F(1,27) = 14.79$; $p < 0.001$] were also found to be reliable. For the task, oral responses were faster than manual responses (286 vs. 545 ms). Regarding the modality, responses were faster in the audiovisual compared to the auditory modality (405 vs. 425 ms). Importantly, a significant group × modality [$F(1,27) = 21.74$; $p < 0.001$] further show that the beneficial effect of audiovisual presentation occurred with noisy stimuli in Group B (461 vs. 507 ms) but not with clear stimuli in Group A (354 vs. 349 ms; see **Figure 2** and **Table 1**).

In sum, the above-mentioned results thus replicate and extend previous studies on speech shadowing (references) by demonstrating a clear advantage of oral responses with both clear and noisy stimuli. In addition, compared to unimodal auditory stimuli, audiovisual stimuli led to faster RTs but only with noisy stimuli. Interestingly, no interaction was found between these two effects thus suggesting they occurred independently.

It should be however, mentioned that these effects also appear dependent on the perceived speech syllable. Overall, significant differences were found between syllables [$F(2,54) = 9.66$; $p < 0.001$], with faster RTs for /apa/ (383 ms) than for /ata/ (438 ms) and /aka/ (424 ms). In addition, a significant syllable × modality interaction was observed [$F(2,54) = 10.88$; $p < 0.001$]. *Post hoc* analyses showed that RTs for /apa/ were faster in the audiovisual compared to the auditory conditions (357 vs. 410 ms), while RTs for /ata/ and /aka/ did not differ in the two modalities (/ata/: 441 vs. 435 ms; /aka/: 418 vs. 430 ms). Finally, a task × modality × syllable interaction was found [$F(2,54) = 6.49$; $p < 0.005$].In the auditory modality, no significant RT differences were observed between syllables for both oral (/apa/: 282 ms; /ata/: 308 ms; /aka/: 312 ms) and manual responses (/apa/: 538 ms; /ata/: 563 ms; /aka/: 549 ms). However, in the audiovisual modality, faster oral RTs occurred for /apa/ compared to /ata/ and /aka/ (/apa/: 237 ms; /ata/: 283 ms; /aka/: 291 ms) while faster manual RTs occurred for

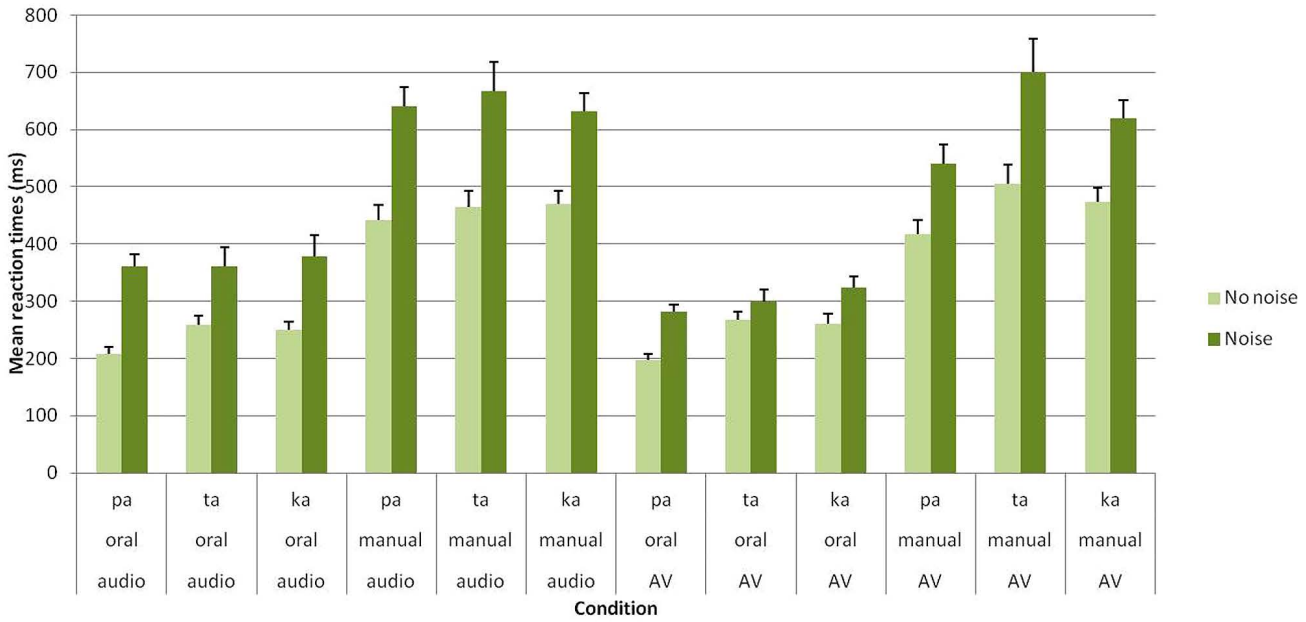


FIGURE 2 | Mean RTs (in ms) of correct identification for /apa/, /ata/, and /aka/ syllables in each group (Group A without noise vs. Group B with noise), task (oral vs. manual response) and modality of presentation (auditory vs. audiovisual). Error bars represent standard errors of the mean.

Table 1 | Significant effects and interactions for all variables.

	Reaction times	Percentage of correct responses
Group effect	$p < 0.001$	$p < 0.001$
Task effect	$p < 0.001$	$p < 0.001$
Modality effect	$p < 0.001$	$p < 0.001$
Group × Modality	$p < 0.001$	$p < 0.001$
Goup × Task	$p < 0.001$	$p < 0.001$
Syllable effect	$p < 0.001$	$p < 0.001$
Syllable × Modality	$p < 0.001$	n.s.
Syllable × Group	n.s.	$p < 0.001$
Syllable × Task	n.s.	$p < 0.001$
Group × Task × Syllable	n.s.	$p < 0.001$
Modality × Task × Syllable	$p < 0.005$	n.s.

/apa/ compared to /aka/ and for /aka/ compared to /ata/ (/apa/: 476 ms; /ata/: 599 ms; /aka/: 544 ms). Taken together, these results likely indicate that visual information processing depends on the level of visual specificity of the presented consonant, with notably a clear advantage for /apa/ syllable (including a bilabial stop consonant). No other effect or interaction were found to be significant.

PERCENTAGE OF CORRECT RESPONSES

The main effect of group was significant [$F(1,27) = 266.28$; $p < 0.001$] with a higher percentage of correct responses for clear stimuli in Group A (95%) than for noisy stimuli in Group B (61%). Importantly, significant main effects were also found for both the task [$F1(1,27) = 69.40$; $p < 0.001$] and the modality [$F(1,27) = 52.39$; $p < 0.001$]. Concerning the task, an important decrease of correct responses was observed for oral compared to manual responses (73 vs. 85%). As indicated by a significant group × task interaction [$F(1,27) = 38.67$; $p < 0.001$], this effect only appeared with noisy stimuli in Group B (71 vs. 50%) while no differences were observed between oral and manual responses with clear stimuli in Group A (93 vs. 98%). For the modality, the audiovisual modality led to higher correct responses than the auditory modality (82 vs. 75%). Importantly, as indicated by a significant group × modality interaction [$F(1,27) = 72.36$; $p < 0.001$], no differences were observed between the two modalities with clear stimuli in Group A, (96 vs. 95%) whereas with noisy stimuli in Group B the audiovisual modality led to higher correct responses (68%) compared to the auditory modality (53%; see **Figure 3** and **Table 1**).

In sum, for noisy stimuli, these results demonstrate a beneficial effect of audiovisual presentation together with a dramatic increase of errors for oral responses. As for RTs, no interaction was however, found between these two effects.

Apart from these results, several other effect and interactions occurred depending on the perceived syllable. First, the main effect of syllable was reliable [$F(2,54) = 25.72$; $p < 0.001$], with a higher recognition of /apa/ (80%) compared to /aka/ (78%) as well as for /aka/ compared to /ata/ (70%). Second, a group × syllable interaction [$F(2,54) = 14.43$; $p < 0.001$] was found. With clear stimuli in Group A, no differences were observed between the three syllables (/apa/: 98% /ata/: 94% /aka/: 94%) while, with noisy stimuli in Group B, /apa/ (77%) was better

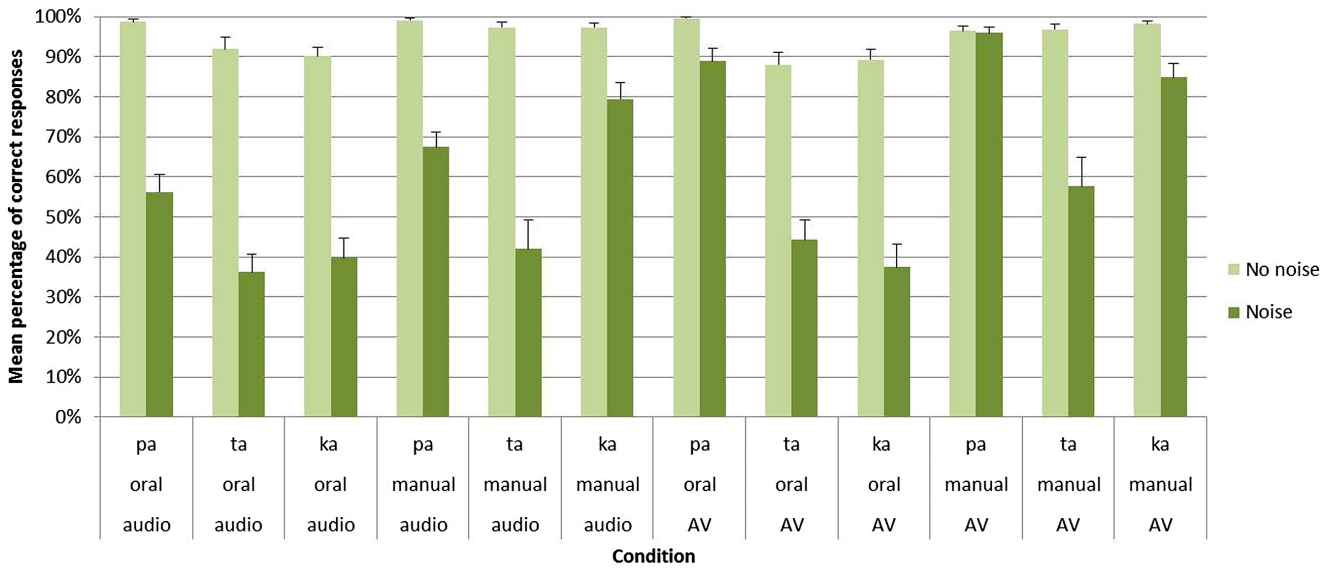


FIGURE 3 | Mean percentage of correct identification for /apa/, /ata/, and /aka/ syllables in each group (Group A without noise vs. Group B with noise), task (oral vs. manual response) and modality of presentation (auditory vs. audiovisual). Error bars represent standard errors of the mean.

recognized than /aka/ (60%) which was itself better recognized than /ata/ (45%). Third, both a task × syllable [$F(2,54) = 22.30$; $p < 0.001$] and a group × task × syllable [$F(2,54) = 11.98$; $p < 0.001$] interactions were observed. With clear stimuli in Group A, no differences were found between the three syllables for both oral (/apa/: 99%; /ata/: 90%; /aka/: 90%) and manual (/apa/: 98%; /ata/: 97%; /aka/: 98%) responses. With noisy stimuli in Group B, /apa/ was better recognized than /ata/ and /aka/ in the oral response mode (/apa/: 73%; /ata/: 40%; /aka/: 39%) while, for manual responses, /apa/ and /aka/ were better recognized than /ata/ (/apa/: 82%; /ata/: 50%; /aka/: 82%). While the three syllables were almost perfectly recognized without noise in Group A, these results demonstrate that for noisy stimuli /pa/ was here the most auditory and visual salient syllable. No other effect, alone or in interaction, were found to be significant.

CORRELATION BETWEEN REACTION TIMES AND PERCENTAGE OF CORRECT RESPONSES

For each of the four condition (i.e., oral or manual responses with audio or AV stimuli), a Pearson's correlation analysis was performed in order to measure the relationship between RTs and percentage of correct responses (with one correlation point computed for each participant and each syllable, irrespective of the group; see **Figure 4**). For all conditions, the higher was the recognition score, the faster was the response; with a negative correlation between RT and response accuracy observed for oral [$r = -0.56$, $t(85) = 17.20$; $p < 0.001$] and manual [$r = -0.41$, $t(85) = 14.32$; $p < 0.001$] responses to audio stimuli as well as for oral [$r = -0.24$, $t(85) = 14.36$; $p < 0.001$] and manual [$r = -0.32$, $t(85) = 9.83$; $p < 0.001$] responses to AV stimuli.

DISCUSSION

We will focus the Discussion on the effects associated with the two major components of our study: the mode of responses, oral vs. manual, and the modality of presentation, auditory vs. audiovisual, and the way they impacted participants' responses.

EFFECT OF TASK: ORAL vs. MANUAL MODE OF RESPONSE

Without noise (Group A), RTs were significantly faster for oral than for manual response (240 vs. 462 ms), with a non-significant decrease in accuracy in the oral response task (93 vs. 98%). RTs in the oral mode are consistent with those found by Fowler et al. (2003; 248 ms) and Porter and Castellanos (1980; 223 ms) in their multiple choice task. Accuracy in the oral mode happens however, to be higher in our study than in Fowler et al. (2003; 86%) and in Porter and Castellanos (1980; 77%) studies. These differences could be due either to the clarity of the provided stimuli or to the sound level at which the presentation was done (the shadowing of the initial vowel leads to a concurrent sound produced by the participant which may hide to a certain extent the perception of the target plosive to identify). The interpretation by both Porter and Castellanos (1980) and Fowler et al. (2003) of the quick response in the oral mode is done in reference to motor theories of speech perception, in which the speech input would be transformed into a motor representation (Liberman and Mattingly, 1985) or would directly be perceived as an orofacial gesture (Fowler and Smith, 1986). This would enable the orofacial system to respond in a highly rapid way, since the percept would already be in the adequate

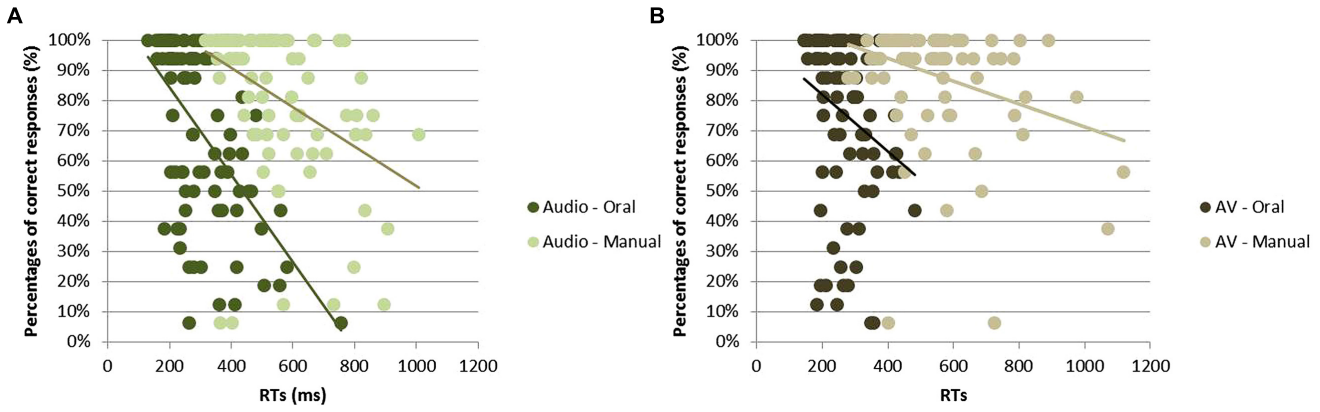


FIGURE 4 | Correlation between RTs (in ms) and correct identification (in %) for all syllables in response to audio stimuli (A) and to audiovisual stimuli (B).

motor format; and more quickly than the manual system which would need a translation stage between decision and response. More generally, these results appear in line with stimulus–response compatibility effects that suggest a common coding in perception and action (for reviews, see Prinz, 1997; Hommel et al., 2001).

However, the observed results with noisy stimuli (Group B) shed a quite new light on this reasoning. Indeed, while RTs stay much faster in close-shadowing (334 vs. 633 ms), accuracy happens to abruptly decrease from the oral to the manual task (50 vs. 71%). This requires modifying the above-mentioned interpretation by Fowler et al. (2003) and Porter and Castellanos (1980) to a certain extent. We will here propose a tentative explanation in the framework of the perceptuo-motor feed forward-feedback model of speech perception proposed by Skipper et al. (2007).

Skipper et al. (2007) propose a speech perception model that they refer to the "analysis-by-synthesis" approach (Halle and Stevens, 1959; Stevens and Halle, 1967; see a review in Bever and Poeppel, 2010). This model involves a processing loop between auditory and motor areas in the human brain (**Figure 3A**). After an initial stage of auditory processing (primary auditory cortex, A1, and further processing in the secondary cortex and associative areas: stage 1 in **Figure 5A**), the auditory cortex would generate a phonemic hypothesis associated with articulatory goals (in the pars opercularis of the inferior frontal gyrus, POp). Then motor commands corresponding to this initial prediction would be stimulated (in the ventral premotor cortex, PMv, and primary motor cortex, M1: stage 2 in **Figure 5A**), leading to the production of an efferent copy sent back to the auditory cortex in order to be compared with the auditory input (stage 3 in **Figure 5A**).

This model could be used as a basis for attempting to interpret our own data (**Figure 5B**). For this aim, we assume that oral and manual responses are generated at two different stages in the processing loop. Oral responses would be generated at stage 2, in line which the assumption by Porter and Castellanos (1980) or Fowler et al. (2003). When the information from the auditory cortex would have been transferred to the POp and generate motor commands in the motor cortex (feedforward strand), the orofacial system, already pre-activated since the beginning of the close-shadowing experiment to allow the participant to answer as quickly as possible, would generate an oral answer produced by these motor commands (stage 2' in **Figure 5B**). This makes the oral answer faster, but it also happens to be inaccurate, which is in line with the proposal by Skipper et al. (2007) that it is only a first hypothesis (possibly rough) that needs to be further refined in a later stage. At stage 2, however, the manual system would not receive specific stimulation enabling it to generate an answer. However, at the next stage (stage 3), the feedback transfer of articulatory information to the auditory cortex, thanks to the efference copy, would provide a more accurate answer that can now be transferred to the manual system for answer (stage 3' in **Figure 5B**). As a consequence, RTs for manual responses would be slower than for oral response, but the responses would be more accurate because, contrary to processing for oral responses, in the manual decision mode, predictions would be confirmed and tuned in the auditory cortex before the final decision would be sent to manual motor commands (pressing the appropriate key).

Of course, this explanation is probably too simple to account for all aspects of our data. The increase in RTs with noisy stimuli (Group B), classical in any categorization experiment, requires some processing expanding over time at various stages in the loop displayed in **Figure 5**. In addition, the fact that the increase is the same in the oral and manual tasks (with no interaction between group and task for RTs) suggests that expansion should basically take place at stages 1 and 2 rather than 3 (but many variants could certainly be suggested). The crucial aspect of our results is that a pure motor translation process typical of motor theories, though compatible with faster RTs in the oral mode, does not appear in line with the associated decrease in response accuracy. On the contrary, it fits well with perceptuo-motor theories of speech perception such as the one proposed by Skipper et al. (2007; see also a computational implementation of a perceptuo-motor theory in Moulin-Frier et al., 2012).

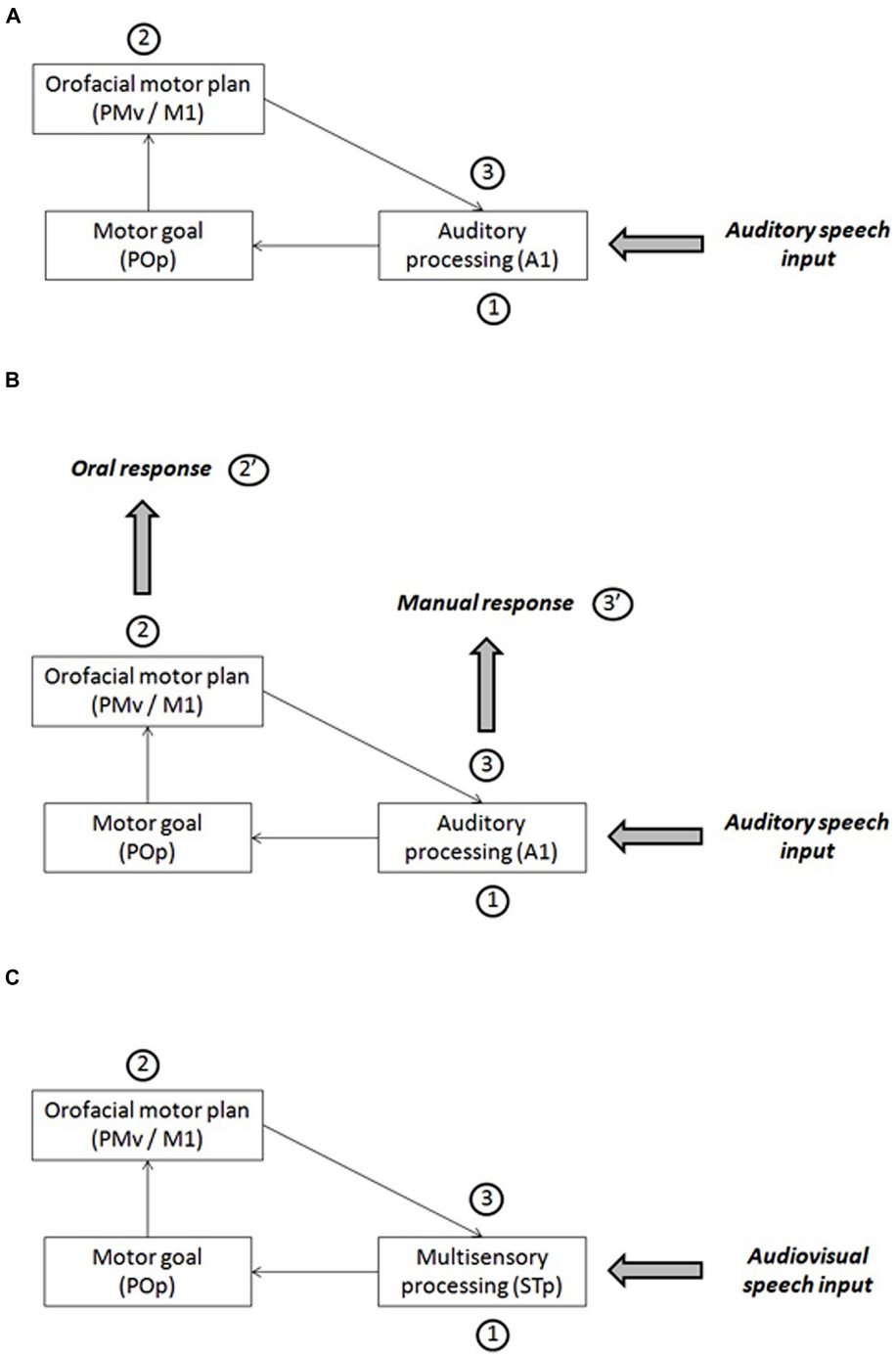


FIGURE 5 | (A) A sketch of Skipper et al.'s (2007) model of speech perception for auditory inputs. **(B)** A possible interpretation of the results about the response mode in the framework of Skipper et al.'s (2007) model. **(C)** A sketch of Skipper et al.'s (2007) model of speech perception for audiovisual inputs.

Partly in line with this hypothesis, it has to be mentioned that close-shadowing as well choral speech are well-known to be a powerful fluency enhancer that is thought to correct deficits in sensorimotor integration (i.e., weak internal modeling; see Harbison et al., 1989; Kalinowski and Saltuklaroglu, 2003).

EFFECTS OF MODALITY: AUDIO vs. AUDIOVISUAL

Effects of modality in our study are only present in the Group B with noisy stimuli. In the auditory modality, RTs are slower than in the audiovisual modality, and proportions of correct answer are lower. Taken together, this shows a clear benefit of adding the visual modality to the auditory input, which is consistent with all

previous studies since Sumby and Pollack (1954) which display an audiovisual benefit to speech recognition in noise conditions. In our study, the audiovisual advantage is present only for the /apa/ syllable which is classical and in line with the higher visibility of the lips movement associated with the bilabial /p/, and the high degree of confusion between visual movements associated with /t/ or /k/ consonants, generally considered to belong to the same visemic class. These effects of modality are not displayed in Group A with clear stimuli. This is probably because in this group, RTs in the auditory modality were already too short and proportions of correct answer were too high to be improved by the visual input (floor effect).

An interesting point is that there is no significant interaction between modality and task that is to say that the decrease of RTs and the increase of proportions of correct responses from the audio to the audiovisual modality are similar for the manual and oral tasks. Once again we will attempt to interpret this lack of interaction to the model proposed by Skipper et al. (2007).

In their model, Skipper et al. (2007) propose that the auditory and visual information, after a preliminary stage of unisensory processing respectively in visual and auditory areas, would converge in the multisensory area STp in the posterior superior temporal cortex (stage 1 in **Figure 5C**). Therefore, in case of multisensory inputs, the first hypothesis would be actually multisensory rather than uniquely auditory. From this basis, here again, a phonemic hypothesis associated with articulatory goals would be generated in POp and evoke motor commands in PMv/M1 (stage 2 in **Figure 5C**), and the efferent copy would produce in STp an auditory prediction to be compared with the auditory input (stage 3 in **Figure 5C**).In our study, audiovisual interactions in stage 1 would refine sensory processing and produce quicker and more accurate phonemic hypotheses in stage 2, which is the stage where, in our interpretation, oral responses would be generated (stage 2' in **Figure 5B**). Then, the same gain in speed and accuracy would be propagated toward stage 3 where manual responses would be generated (stage 3' in **Figure 5B**). Therefore, there is no strong reason to expect differences in visual gain between oral and manual tasks, the gain being essentially determined as soon as stage 1 in the model.

In summary, the results of the present study suggest that oral and manual responses are generated at two different stages in the whole perceptual chain. In the framework of an "analysis-by-synthesis" approach, manual responses would be provided only at the end of the entire loop, following motor predictions then commands themselves generating a multisensory hypothesis compared to the incident multisensory stream. However, oral responses would be produced at an earlier stage where motor commands are generated, causing faster but less precise responses. The visual input would increase speed and accuracy for sufficiently visible phonemes (e.g., /p/) in case of adverse listening conditions (such as noise). Once again, it is important to stress that other interpretations or frameworks could be provided. But globally, we argue that the whole set of results of this study seems to require a perceptuo-motor theory of speech perception in which the auditory and visual streams are integrated together and with internally generated motor representations before a final decision may be available.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency (ANR) through funding for the PlasModyproject (Plasticity and Multimodality in Oral Communication for the Deaf).

REFERENCES

Bever, T. G., and Poeppel, D. (2010). Analysis by synthesis: a (re-)emerging program of research for language and vision. *Biolinguistics* 4, 174–200.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., and Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nat. Neurosci.* 7, 295–301. doi: 10.1038/nn1198

Boersma, P., and Weenink, D. (2013). *Praat: Doing Phonetics by Computer. Computer Program, Version 5.3.42.* Available at: http://www.praat.org/ [accessed March 2, 2013].

Callan, D., Jones, J., Callan, A., and Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *Neuroimage* 22, 1182–1194. doi: 10.1016/j.neuroimage.2004.03.006

Cotton, J. C. (1935). Normal 'visual hearing.' *Science* 82, 592–593. doi: 10.1126/science.82.2138.592

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex* 48, 882–887. doi: 10.1016/j.cortex.2011.05.017

D'Ausilio, A., Jarmolowska, J., Busan, P., Bufalari, I., and Craighero, L. (2011). Tongue corticospinal modulation during attended verbal stimuli: priming and coarticulation effects. *Neuropsychologia* 49, 3670–3676. doi: 10.1016/j.neuropsychologia.2011.09.022

D'Ausilio, A., Pulvermuller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

Davis, C., and Kim, J. (2001). Repeating and remembering foreign language words: implications for language teaching system. *Artif. Intell. Rev.* 16, 37–47. doi: 10.1023/A:1011086120667

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.* 55, 149–179. doi: 10.1146/annurev.psych.55.090902.142028

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.* 15, 399–402. doi: 10.1046/j.0953-816x.2001.01874.x

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct realistic perspective. *J. Phon.* 14, 3–28.

Fowler, C. A., Brown, J. M., Sabadini, L., and Weihing, J. (2003). Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *J. Mem. Lang.* 49, 296–314. doi: 10.1016/S0749-596X(03)00072-X

Fowler, C. A., and Smith, M. (1986). "Speech perception as "vector analysis": an approach to the problems of segmentation and invariance," in *Invariance and Variability of Speech Processes*, eds J. Perkell and D. Klatt (Hillsdale, NJ: Lawrence Earlbaum Associates), 123–136.

Galantucci, B., Fowler, C. A., and Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychon. Bull. Rev.* 13, 361–377. doi: 10.3758/BF03193857

Grabski, K., Schwartz, J. L., Lamalle, L., Vilain, C., Vallée, N., Baciu, M., et al. (2013a). Shared and distinct neural correlates of vowel perception and production. *J. Neurolinguistics* 26, 384–408. doi: 10.1016/j.jneuroling.2012.11.003

Grabski, K., Tremblay, P., Gracco, V. L., Girin, L., and Sato, M. (2013b). A mediating role of the auditory dorsal pathway in selective adaptation to speech: a state-dependent transcranial magnetic stimulation study. *Brain Res.* 1515, 55–65. doi: 10.1016/j.brainres.2013.03.024

Halle, M., and Stevens, K. N. (1959). "Analysis by synthesis," in *Proceedings of the Seminar on Speech Compression and Processing L.G. Hanscom Field, Bedford, Massachusetts*, eds W. Wathen-Dunn and L. E. Woods (Bedford, MA: USAF Cambridge Research Center).

Harbison, D. C. Jr., Porter, R. J. Jr., and Tobey, E. A. (1989). Shadowed and simple reaction times in stutterers and nonstutterers. *J. Acoust. Soc. Am.* 86, 1277–1284. doi: 10.1121/1.398742

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Hommel, B., Musseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* 24, 849–878. doi: 10.1017/S0140525X01000103

Kalinowski, J., and Saltuklaroglu, T. (2003). Choral speech: the amelioration of stuttering via imitation and the mirror neuronal system. *Neurosci. Biobehav. Rev.* 27, 339–347. doi: 10.1016/S0149-7634(03)00063-0

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the HandH theory", in *Speech Production and Speech Modelling*, eds W. J. Hardcastle and A. Marchal (Dordrecht: Kluwer Academic Publishers), 403–439. doi: 10.1007/978-94-009-2037-8_16

Lindblom, B., and Maddieson, I. (1988). "Phonetic universals in consonant systems," in *Language, Speech and Mind*, eds C. Li and L. M. Hyman (London: Routledge), 62–78.

Lotto, A. J., Hickok, G. S., and Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends Cogn. Sci.* 13, 110–114. doi: 10.1016/j.tics.2008.11.008

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization.* New York: Oxford University Press.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Curr. Biol.* 17, 1692–1696. doi: 10.1016/j.cub.2007.08.064

Möttönen, R., Järveläinen, J., Sams, M., and Hari, R. (2004). Viewing speech modulates activity in the left SI mouth cortex. *Neuroimage* 24, 731–737. doi: 10.1016/j.neuroimage.2004.10.011

Möttönen, R., and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *J. Neurosci.* 29, 9819–9825. doi: 10.1523/JNEUROSCI.6018-08.2009

Möttönen, R., and Watkins, K. E. (2012). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology* 26, 1103–1118. doi: 10.1080/02687038.2011.619515

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., and Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: an exploratory Bayesian modeling study. *Lang. Cogn. Process.* 27, 1240–1263. doi: 10.1080/01690965.2011.645313

Porter, R., and Castellanos, F. (1980). Speech production measures of speech perception: rapid shadowing of VCV syllables. *J. Acoust. Soc. Am.* 67, 1349–1356. doi: 10.1121/1.384187

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154. doi: 10.1080/713752551

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear to understand: a lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Cambell (London: Lawrence Erlbaum), 97–113.

Sato, M., Tremblay, P., and Gracco, V. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang.* 111, 1–7. doi: 10.1016/j.bandl.2009.03.002

Schwartz, J. L., Basirat, A., Ménard, L., and Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguistics* 25, 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Scott, S. K., Rosen, S., Beaman, C. P., Davis, J., and Wise, R. J. S. (2009). The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes. *J. Acoust. Soc. Am.* 125, 1737–1743. doi: 10.1121/1.3050255

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Stevens, K. N., and Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.* 64, 1358–1368. doi: 10.1121/1.382102

Stevens, K. N., and Halle, M. (1967). "Remarks on analysis by synthesis and distinctive features," in *Models for the Perception of Speech and Visual Form*, ed. W. Wathen-Dunn (Cambridge, MA: MIT Press).

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Wilson, S. M., and Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage* 33, 316–325. doi: 10.1016/j.neuroimage.2006.05.032

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., and Schoonhoven, R. (2006). Top-down and bottom-up processes in speech comprehension. *Neuroimage* 32, 1826–1836. doi: 10.1016/j.neuroimage.2006.04.199

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 February 2014; accepted: 22 May 2014; published online: 24 June 2014.
Citation: Scarbel L, Beautemps D, Schwartz J-L and Sato M (2014) The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing. Front. Psychol. ***5****:568. doi: 10.3389/fpsyg.2014.00568*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 April 2014
doi: 10.3389/fpsyg.2014.00323

Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults

***Kaoru Sekiyama*[1,2]*, *Takahiro Soshi*[1] *and Shinichi Sakamoto*[3]**

[1] *Division of Cognitive Psychology, Faculty of Letters, Kumamoto University, Kumamoto, Japan*
[2] *Division of Cognitive Psychology, School of Systems Information Science, Future University, Hakodate, Japan*
[3] *Otodesigners Co., Ltd., Wako, Japan*

Edited by:
Kaisa Tiippana, University of Helsinki, Finland

Reviewed by:
Annalisa Setti, University College Cork, Ireland
Susan Rvachew, McGill University, Canada

***Correspondence:**
Kaoru Sekiyama, Division of Cognitive Psychology, Faculty of Letters, Kumamoto University, 2-40-1, Kurokami, Kumamoto 860-8555, Japan
e-mail: sekiyama@kumamoto-u.ac.jp

Two experiments compared young and older adults in order to examine whether aging leads to a larger dependence on visual articulatory movements in auditory-visual speech perception. These experiments examined accuracy and response time in syllable identification for auditory-visual (AV) congruent and incongruent stimuli. There were also auditory-only (AO) and visual-only (VO) presentation modes. Data were analyzed only for participants with normal hearing. It was found that the older adults were more strongly influenced by visual speech than the younger ones for acoustically identical signal-to-noise ratios (SNRs) of auditory speech (Experiment 1). This was also confirmed when the SNRs of auditory speech were calibrated for the equivalent AO accuracy between the two age groups (Experiment 2). There were no aging-related differences in VO lipreading accuracy. Combined with response time data, this enhanced visual influence for the older adults was likely to be associated with an aging-related delay in auditory processing.

Keywords: speech perception, aging, McGurk effect, response time, hearing level, lipreading, auditory-visual integration

INTRODUCTION

In face-to-face speech communication, perceivers use not only auditory speech information, but also visual articulatory information from the talker's face (Stork and Hennecke, 1996; Campbell et al., 1998; Massaro, 1998; Bailly et al., 2012). The use of visual information is especially prominent when auditory speech is degraded (Sumby and Pollack, 1954; Grant and Seitz, 2000; Schwartz et al., 2004; Ross et al., 2007). The contribution of visual speech to undegraded auditory speech is easily demonstrated when participants are presented with incongruent visual speech, as in the McGurk effect (McGurk and MacDonald, 1976). In the original report by McGurk and MacDonald (1976), auditory syllables with relatively high intelligibility (94% unisensory accuracy on average) were mostly perceived as auditorily wrong syllables when presented with incongruent visual speech. For example, auditory /ba/ stimuli were mostly perceived as "da" when presented with visual /ga/, indicating perceptual fusion of auditory and visual speech. Thus, the McGurk effect paradigm is a useful tool to measure the visual contribution to intelligible auditory speech.

By using this effect, it has been found that the extent of visual information use, i.e., the size of the McGurk effect varies among different populations (see Schwartz, 2010, for a review). For example, people with cochlear implants show a larger McGurk effect than people with normal hearing (Schorr et al., 2005; Rouger et al., 2008). This finding indicates that the cochlear implant users compensate for hearing impairment by heightened use of visual information. The opposite case has been found among young children in normal hearing populations. Young children show a smaller McGurk effect than adults (McGurk and MacDonald, 1976; Massaro et al., 1986; Tremblay et al., 2007; Sekiyama and Burnham, 2008). The greater reliance on auditory speech is perhaps largely due to the poorer lipreading ability of children (Massaro et al., 1986; Sekiyama and Burnham, 2008; Chen and Hazan, 2009).

The group differences in the above examples can be largely accounted for by the accuracy or confusability of unisensory information. That is, the sensory modality with less confusion plays a larger role, resulting in optimal integration as expressed by maximum-likelihood estimation or a Bayesian model (Massaro, 1987, 1998; for an improved version of the Bayesian model, see Schwartz, 2010; For a different approach, see also Braida, 1991; Grant et al., 1998). However, in some cases it is difficult to explain group differences by the unisensory accuracy alone. Language background could be one such case. For example, adult native speakers of Japanese show a smaller McGurk effect, and so a stronger auditory dependence, compared with English native speakers (Sekiyama and Tohkura, 1991, 1993; Kuhl et al., 1994; Sekiyama, 1994; Sekiyama and Burnham, 2008; also see ANOVA results of Massaro et al., 1993). Although some of these language differences may be accounted for by unisensory accuracy to some extent (Massaro et al., 1993), the Japanese-English differences in the McGurk effect could be observed when unisensory accuracy was equivalent between the two groups for both auditory and visual speech (Sekiyama and Burnham, 2008). Such a case suggests another factor affecting auditory-visual integration.

Recent neuro-cognitive studies have revealed that the integration of auditory and visual information is facilitated if the two information streams in the brain converge during an optimal time window (Van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Schroeder et al., 2008; also see Altieri et al., 2011, for a review). Considering the importance of such

temporal characteristics, response time data can provide some insight into the integration processes in various types of populations. Sekiyama and Burnham (2008) compared unisensory response times (RTs) of Japanese and English-language children and adults, and it was found that English-language adults were faster in visual-only (VO) syllable identification (lipreading) than auditory-only (AO) identification (hearing), whereas Japanese adults' RTs were equivalent for the two conditions. Such group differences in AO-vs.-VO RT could account for language differences in the size of the McGurk effect in the auditory-visual (AV) condition. The AO − VO RT difference was not found in 6-year-olds in either of the two language groups and the McGurk effect was generally weak at this age. Based on these results, a "visual priming hypothesis" was proposed whereby the visual contribution is larger when an individual processes visual speech faster than auditory speech (Sekiyama and Burnham, 2008).

The present study investigated how aging affects auditory-visual speech perception by comparing the McGurk effect in young and older adults. In order to do this, accuracy and speed (RTs) in unisensory speech perception were both examined. Recent studies using event-related potentials (ERPs) have shown that older adults show delays in auditory processing compared with younger adults for both speech (Tremblay and Ross, 2007) and non-speech (Schroeder et al., 1995). Such neurophysiological temporal characteristics lead to an assumption that older adults have a greater visual priming effect than young adults due to delayed auditory processing. In fact, this is indirectly suggested by a previous ERP study on auditory-visual speech perception showing that multisensory temporal facilitation was greater for older adults than young adults in perceiving congruent auditory-visual spoken words (Winneke and Phillips, 2011). However, precise examinations are still necessary by measuring the RTs for each of the AO, VO, and AV conditions. Moreover, by using the McGurk effect paradigm, it is possible to investigate the relationship between temporal characteristics and how visual information is incorporated in perceived speech.

As for the susceptibility of the McGurk effect, we predicted that older adults would yield a larger McGurk effect than younger adults based on the above-mentioned delay in auditory processing. The delay may be associated with the well-documented hearing threshold decline in older adults (e.g., Glorig and Nixon, 1962; CHABA, Committee on Hearing, Bioacoustics, and Biomechanics, 1988; Pichora-Fuller and MacDonald, 2009). Combined with the fact that visual contribution is generally larger in harder hearing circumstances (Sumby and Pollack, 1954; Grant and Seitz, 2000; Schwartz et al., 2004; Ross et al., 2007), it is thought that older adults tend to utilize visual information more to compensate for their declined hearing. Such a prediction is supported by older adults' greater attention allocation to a speaker's mouth compared with younger adults (Thompson and Malloy, 2004). Also, a few studies actually suggested an aging-related increase in the McGurk effect (Thompson, 1995; Behne et al., 2007; Setti et al., 2013). However, other highly controlled studies have reported non-significant differences between young and older adults in auditory-visual speech perception (Cienkowski and Carney, 2002; Sommers et al., 2005). Cienkowski and Carney (2002) found that the aging-related difference in the size of the McGurk effect was not clear when confined to participants with age-appropriate hearing levels. They presented AV-incongruent McGurk stimuli (e.g., auditory /bi/ with visual /gi/) and older adults were compared with young controls whose auditory thresholds were shifted with noise to match the older adults. The results included some response pattern differences between groups depending on the talker and consonant, but on average, older adults integrated auditory and visual information as much as young controls. Likewise, Sommers et al. (2005) presented congruent auditory-visual speech (consonants, words, and sentences) to normal hearing older and young adults. Each participant was tested at a customized signal-to-noise ratio (SNR) to equate auditory intelligibility across individuals. The results showed the same degree of auditory-visual integration for the two groups, after factoring out lipreading performance differences. Otherwise, older adults appeared to benefit from visual speech less than younger adults.

As this statistical procedure by Sommers et al. (2005) highlights, lipreading performance was poorer in older adults compared with younger adults in the above two studies (Cienkowski and Carney, 2002; Sommers et al., 2005). This makes it complicated to compare the two age groups in terms of auditory-visual integration. The literature shows that the poorer lipreading performance of older adults depends on age and the speech material (Shoop and Binnie, 1979; Walden et al., 1993). For example, Shoop and Binnie (1979) found that aging-related decline of lipreading accuracy was observed for sentences starting from 40 years, but for consonants in a consonant + /a/ context, the decline was not very evident until 70 years. The above two studies included participants over 70 years (age range 65–74 years in Cienkowski and Carney, 2002; Mean = 70.2 and $SD = 6.8$ in Sommers et al., 2005); therefore, lowering the age limit may help reduce differences in lipreading performance between older and young groups. Consequently, this study limited older participants to the "young-old," with an age limit up to 65 years. Also, our speech materials were consonants in a consonant + vowel /a/ context as in Shoop and Binnie (1979).

In addition, some controls were necessary over the auditory dimension to deal with age-related differences in hearing thresholds. Aging-related decline in hearing thresholds starts in the early thirties, and a significant decline occurs before the age of 65 (CHABA, Committee on Hearing, Bioacoustics, and Biomechanics, 1988). To control such age-related differences in hearing thresholds, auditory noise is often used by differing SNRs between age groups. Previous studies compared AV speech perception between older and young adults either with the same SNRs (Thompson, 1995; Behne et al., 2007; Setti et al., 2013) or calibrated SNRs between groups (Cienkowski and Carney, 2002; Sommers et al., 2005). Only in the former were age-related differences in AV performances found. Of course, the calibration of SNRs is important to investigate different groups with different hearing thresholds; however, calibrating them is not so simple. In Cienkowski and Carney (2002), the young control group received band-pass noise, which resulted in poorer AO performance in the control group compared with an older group who were not given the noise. In Sommers et al. (2005), individually customized SNRs were used for each participant, but their SNRs for 50% correct

AO performance level may be too low for our McGurk effect paradigm. Considering these facts, we took two approaches. In Experiment 1, young and older adults were compared under the same auditory SNRs. In Experiment 2, the two age groups were tested under calibrated SNRs (estimated from the group results of Experiment 1 for AO perceptual equivalence). Based on the previous research, we predicted an aging-related increase in the McGurk effect in Experiment 1. If an aging-related increase is also observed in Experiment 2, it would be a novel finding.

The purpose of the present study was two-fold. First, to examine whether or not young-olds with normal hearing use visual information more than young adults. If so, the second purpose was to test the visual priming hypothesis from a previous study (Sekiyama and Burnham, 2008). The hypothesis postulates that the visual contribution will be large for those who process visual speech faster than auditory speech compared with those who process visual and auditory speech at about the same speed. We investigated whether older adults are more likely to show a greater visual priming effect in auditory-visual speech perception than younger adults. Thus, we focused on the RT differences between the AO and VO conditions as the basis for the visual priming effect. We predicted that the AO − VO RT difference would be larger for the older adults based on the delayed auditory processing reported in ERP studies (Tremblay and Ross, 2007). With the perceptually equivalent SNRs in Experiment 2, we tested whether or not older adults were still more visually influenced when unisensory auditory accuracy was the same across the two age groups. In addition, the equivalent auditory accuracy guaranteed that differences in RT represent differences in processing speed, at least for the AO condition. Before testing the hypothesis in Experiment 2 with calibrated SNRs, Experiment 1 was conducted to determine how to calibrate SNRs to obtain equivalent auditory accuracy between the younger and older adults.

EXPERIMENT 1: SPEECH PERCEPTION PERFORMANCE UNDER THE PHYSICALLY CONTROLLED SNRs

The purposes of Experiment 1 were (1) to describe age-related differences in auditory-visual speech perception under various auditory SNRs which were physically the same for the older and younger groups, and (2) to determine SNRs for each age group under which AO accuracy was equivalent between the two age groups.

MATERIALS AND METHODS

Participants

Thirty-four Japanese monolingual speakers participated in the experiment. The experimental protocol was approved by the Research Ethics Committee at Future University Hakodate, and all the participants filled written consent form before the experiment. Sixteen older participants (8 males, 8 females) were recruited through the City Employment Agency for Older People Hakodate. They were aged between 60 and 65 years old, and were recruited after reporting normal hearing on a self-reported basis. These people were still actively working after retirement doing part-time jobs through the Agency. Eighteen younger participants (10 males, 8 females) were university students aged between 19 and 21 years old. All of the older and younger participants had normal or corrected-to-normal vision. The experimental data were analyzed after screening the participants by hearing threshold (measured by an audiometer: Rion AA-73). Along with the criterion defined by the World Health Organization, the threshold was set to a ≤25 dB hearing level (HL) of averaged HLs of 500, 1000, 2000, and 4000 Hz. Twelve older participants met the threshold criterion (Mean ± *SD* : 18.0 ± 3.3 dB HL), while four older participants failed to meet the criterion (30.2 ± 0.8 dB HL), so were excluded from the analysis. All of the younger participants met the criterion (6.6 ± 3.4 dB HL), and were included in the analysis. The ages in the final sample were as follows: older (Mean = 62.3, *SD* = 1.8 years), younger (Mean = 20.4, *SD* = 0.9 years).

Stimuli

The stimuli consisted of /ba/, /da/, and /ga/ uttered by three talkers (two male and one female, native Japanese speakers). The utterances were videotaped, digitized, and edited on computer to produce AO, VO, and AV stimuli. Video digitizing was done at 29.97 frames/s in 640 × 480 pixels, and audio digitizing at 32000 Hz in 16 bit; each stimulus was created as a 2300 ms movie of a monosyllabic utterance. The duration of acoustic speech signals in each movie was approximately 290 ms on average. The movie file was edited with frame unit accuracy (33.3 ms), and the sound portion was additionally edited with 1 ms accuracy so that the sound onset was at 900 ms for each movie clip (for more details, see Sekiyama et al., 2003). Half of the AV stimuli were congruent (AVc condition: e.g., auditory /ba/ and visual /ba/, i.e., AbVb). The other half of the AV stimuli were so-called McGurk-type incongruents (AVi condition): that is, the auditory part (e.g., /ba/) is incongruent with the visual articulation (e.g., /ga/). Three kinds of McGurk-type stimuli were created by combining within-talker auditory and visual components (AbVg, AdVb, AgVb). The VO stimuli, one each for /ba/, /da/, and /ga/, were created by cutting out the audio track. In the AO stimuli, one each for /ba/, /da/, and /ga/, the video of a talking face was replaced by the still face of the talker with the mouth neutrally closed. In total, there were 9 AO stimuli (3 consonants × 3 talkers), 9 VO stimuli (3 consonants × 3 talkers), and 18 AV stimuli (3 auditory consonants × 3 talkers × 2 AV-congruent (AVc) /incongruent (AVi) types).

Auditory intelligibility was manipulated for four levels of auditory intelligibility by adding band noise (300–12000 Hz) with SNRs of 0, +6, +12, and +18 dB. The speech was always presented at 65 dB sound pressure level (SPL) and the noise level was varied. There was no noise-free condition because the previous results indicated that SNRs higher than +12 dB would result in the same performances as for a noise-free condition, at least for the younger participants (Sekiyama and Burnham, 2008).

Procedure

Each participant was tested individually in a sound-attenuated room. The stimuli were presented from a personal computer onto a 17-inch CRT monitor and through a loudspeaker using in-house software. Experimental conditions were blocked depending on the presentation mode (AV, AO, VO) and the SNR of the auditory stimuli (0, +6, +12, +18 dB), and there were two repetitions

of each stimulus in a block (2 × 9 stimuli = 18 trials in each block in the AO and VO conditions, and 2 × 18 = 36 trials for each block in the AV conditions). Each participant was given the AV condition first. Half of the participants were tested with an AV-AO-VO order, and the other half with an AV-VO-AO order. In the AV and AO conditions, the speech was presented at 65 dB SPL at the participant's ear level, and the SNRs, 0, +6, +12, and +18 dB were determined by the intensity of the added band noise. The SNR varied across blocks in an increasing manner for half of the participants, and in a decreasing manner for the remaining participants.

Within each block, the stimuli were presented in random order. The participants were asked to watch and listen to each stimulus, decide what they perceived, and press one of three buttons for a "ba," "da," or "ga" response accurately and without delay. After each movie file was played, the last frame remained on the screen until one of the three buttons was pressed. Responses were made on a game controller, with input to the computer such that the responses were stored. The onset of the next stimulus was 1500 ms after the button press.

Before starting the first block of each of the AV, AO, and VO conditions, practice trials were given for nine, six, and six times, respectively, using stimuli not used in the test trials. Excluding these practice trials, the total number of trials per participant was 234 (18 trials × 4 SNRs for AO, 18 trials for VO, and 36 trials × 4 SNRs for the AV conditions). The experiment took an average of 20 min per participant for the younger group, and 30 min for the older group.

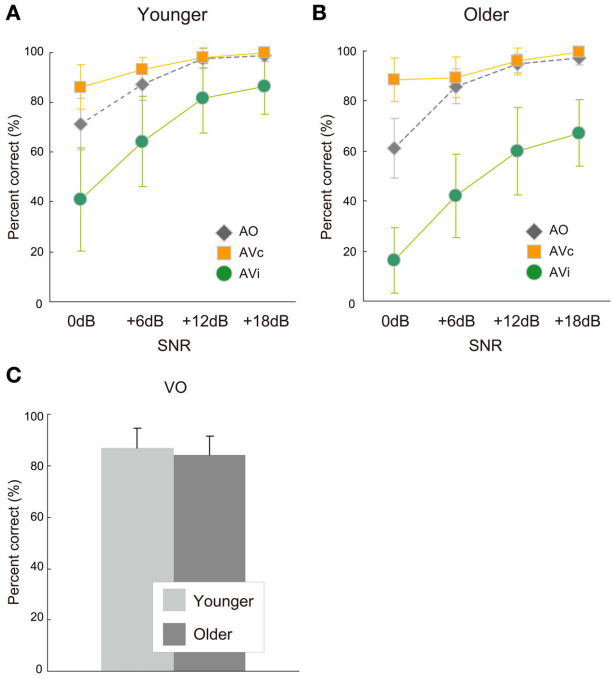


FIGURE 1 | Response accuracy of auditory-only (AO) and auditory-visual (AV) conditions in the (A) younger and (B) older groups and of (C) the visual-only (VO) condition in the younger and older groups. Response accuracy scores are plotted for the four signal-to-noise ratios (SNRs) in the AV and AO conditions. AV conditions included two conditions: AV congruent (AVc) or incongruent (AVi) conditions.

Statistical analysis

Statistical tests mainly focused on group-related effects; therefore, the effects of SNRs were tested only as an interaction with the age group. Analysis of variance (ANOVA) was conducted with the factors of age group (younger, older) and SNR (0, +6, +12, +18 dB) for percent correct in the AO condition, and visual influence score (AVc − AVi). An unpaired *t*-test was performed for the VO condition to examine group differences. Before each ANOVA, arcsine transformation was conducted on response accuracy to stabilize variance ($Y = 2\text{arcsine}\sqrt{p}$; p: proportion correct) (Howell, 1997). As a result, the visual influence scores were actually (arcAVc − arcAVi). When group-related effects were significant, planned group comparisons were always conducted for each SNR to examine in more detail the group effects. Greenhouse-Geisser correction was performed when the sphericity assumption about the variance of differences was violated, and this was reported with unmodified degrees of freedom and epsilon (ε).

RESULTS

Percent correct responses as a function of the SNR in the AO and AV conditions are shown in **Figure 1A** for the younger group and **Figure 1B** for the older group. **Table 1** indicates mean response accuracy and statistical results in group comparisons. **Figure 1C** compares percent correct responses in the VO condition between the two groups. The correct responses were defined in terms of the auditory component of a stimulus for the AVc and AVi conditions. As described below, the older group was lower in terms of response accuracy in the AO, but not the VO condition, and yielded a larger McGurk effect in the AVi condition compared with the younger group.

The ANOVA for the AO condition showed a significant main effect of age group [$F_{(1,\ 28)} = 11.696$, $p = 0.002$, $\eta^2 = 0.024$], while the age group × SNR interaction was not significant [$F_{(3,\ 84)} = 0.919$, $p = 0.436$, $\eta^2 = 0.005$] (**Figures 1A,B**). Planned comparisons between groups were conducted for each SNR, and significant differences appeared at SNRs of 0 dB and +12 dB [Bonferroni: 0 dB, $p = 0.019$; +6 dB: $p = 0.532$; +12 dB: $p = 0.04$; +18 dB: $p = 0.122$]. These results indicate that the older participants were less accurate in auditory syllable identification under some (low and high) SNR conditions as compared with the younger participants.

For the VO condition, on the other hand, the older and younger groups were not significantly different [$t_{(28)} = 1.247$, $p < 0.223$, Cohen's $d = 0.480$] (**Figure 1C**). This indicates that lipreading performance was not different between the two age groups.

The visual influence scores (AVc − AVi) are shown in **Figure 2**. The ANOVA for this score found a significant main effect of age group, while the age group × SNR interaction was not significant [age group: $F_{(1,\ 28)} = 14.164$, $p = 0.001$, $\eta^2 = 0.188$; age group × SNR: $F_{(3,\ 84)} = 1.266$, $p = 0.291$, $\eta^2 = 0.013$]. Planned comparisons between groups for each SNR confirmed that the older group was more affected by visual information than the

Table 1 | Mean response accuracy (%) in Experiment 1 for younger and older groups.

SNR (dB)	Younger ($n = 18$)					Older ($n = 12$)					Group difference (p-value)		
	AO	AVc	AVi	AVc − AVi	VO	AO	AVc	AVi	AVc − AVi	VO	AO	AVc − AVi	VO
0	71	86	41	45	87	61	88	16	72	84	0.019*	0.003**	0.223
+6	87	93	64	29		86	89	42	47		0.532	0.028*	
+12	98	98	81	16		95	96	60	36		0.04*	0.031*	
+18	99	100	86	14		97	100	67	32		0.122	0.002**	

*SNR, signal-to-noise ratio; AO, auditory-only; AVc, auditory-visual congruent; AVi, auditory-visual incongruent; VO, visual-only; *$p < 0.05$; **$p < 0.01$.*

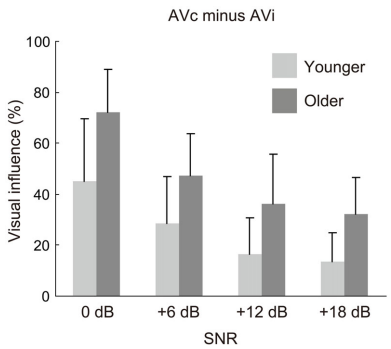


FIGURE 2 | Visual influence scores for the four signal-to-noise ratio (SNR) conditions in the younger and older groups. The scores were calculated by subtraction of percent correct for the auditory-visual incongruent (AVi) condition from the percent correct for the AV-congruent (AVc) condition.

younger at each SNR (Bonferroni: all $p < 0.04$), indicating their general tendency toward greater use of visual information.

DISCUSSION

In Experiment 1, the AV and AO conditions were conducted for four levels of SNRs (0dB, +6dB, +12dB, +18dB), meaning that both age groups were tested under the same physical conditions. Under these experimental settings, the older group showed a larger visual influence than the younger group. For unisensory accuracy, the older group was less accurate in the AO condition than the younger group, while no significant group difference was found in VO accuracy. Taken together, the larger visual influence in the older group might be attributable to their lower AO accuracy. This is not surprising because the older group had a higher hearing threshold on average compared with the younger group, although the thresholds for both groups were within the normal hearing range. Thus, these results are basically in line with the optimal integration model.

While an aging-related increase in visual influence was observed for all four SNRs, the aging-related accuracy degradation in the AO condition was limited to high (+12 dB) and low (0 dB) SNRs. Thus, the relationship between the AO intelligibility and the visual influence was not straightforward in this experiment. To further examine the aging effect in auditory-visual speech perception, it was crucial to investigate whether the group difference in visual influence still existed when AO accuracy was equivalent between the two age groups. To do so, the results from Experiment 1 were used to determine how the AO accuracy should be equated by calibrating SNRs.

From the curves in **Figures 1A,B**, SNRs needed to obtain the same AO performance could be estimated: at a 90% AO accuracy, for example, the younger group's SNR was about 7 dB and the older group's about 11 dB. This was also true when we estimated each individual's SNR point for 90% AO accuracy using an interpolating method and then averaging the estimated SNRs. This group difference was used in Experiment 2 to calibrate SNRs. Perceptually-equivalent SNRs are useful to examine the visual priming hypothesis because measuring RTs should ideally be conducted under a constant accuracy (Luce, 1986).

EXPERIMENT 2: SPEECH PERCEPTION PERFORMANCE UNDER THE PERCEPTUALLY CONTROLLED SNRs

The purposes of Experiment 2 were (1) to examine whether or not an aging-related increase in visual influence could be observed under calibrated auditory SNRs which would result in an equivalent AO accuracy for the older and younger groups, and (2) to investigate age-related changes in the visual precedence time (AO-vs.-VO in RT) to assess our visual priming hypothesis. The results of Experiment 1 revealed that mean SNRs for 90% AO accuracy were 7 dB for the younger group and 11 dB for the older group. We, therefore, set the SNRs here such that SNRs for the older group were 4 dB higher than those for the younger group. In addition, three levels of SNRs were set such that the physical SNRs cover the SNR range in Experiment 1 (0 to +18 dB) because significant group differences in the visual influence (AVc − AVi in percent correct) were observed in all SNRs in Experiment 1. As a result, Experiment 2 used following SNRs for the younger and older participants, respectively: Low (−3, +1 dB), Middle (+7, +11 dB), and High (+17, +21 dB).

METHODS

Participants

Fifty-one Japanese monolingual speakers participated in the experiment. The experimental protocol was approved by the Research Ethics Committee at Future University Hakodate. All of the participants completed a written consent form before the experiment. Participants were similarly recruited as in Experiment 1. Twenty-four older participants (12 males, 12 females) were aged between 60 and 65 years old. Twenty-seven younger participants (14 males, 13 females) were aged between

18 and 21 years old. All of the participants had normal or corrected-to-normal vision. Hearing tests for pure tones were conducted as in Experiment 1. The exclusion criterion for hearing threshold was the same as in Experiment 1. Seventeen older participants met the threshold criterion (16.6 ± 4.3 dB HL), while seven older participants did not (32.3 ± 4.2 dB HL) and were excluded from the analysis. All of the younger participants met the criterion (6.5 ± 3.8 dB HL). The ages of the final sample were as follows: older (Mean = 62.5, *SD* = 1.9 years), younger (Mean = 19.8, *SD* = 1.8 years).

Stimuli

The same speech stimuli as in Experiment 1 were used. In contrast to the same physical SNRs for the two age groups, the present experiment used perceptually equivalent SNRs for the two groups. There were three levels of SNRs (low, middle, high) for each age group. The band noise (300–12000 Hz) was always presented at 54 dB SPL and the speech level was varied so that the SNRs were +1 dB (low), +11 dB (middle), and +21 dB (high) for the older group. Similarly, the speech level was varied for the younger group so that the SNRs were −3 dB (low), +7 dB (middle), and +17 dB (high). Such SNR setting was determined based on the results of Experiment 1, indicating that older participants should be presented with speech louder by 4 dB to obtain the equivalent AO accuracy as the younger participants (see discussion of Experiment 1 and introduction of Experiment 2).

Procedure

The procedure was almost identical to that of Experiment 1. The only difference was that there were two kinds of VO conditions in this experiment. In addition to the VO condition with three-alternative forced choices among "ba," "da," and "ga" (VO3 condition), there was also a two-alternative forced choice condition (VO2 condition) in which the same three visual stimuli were presented for identification of either "ba" or "non-ba." The VO2 condition was introduced based on a pilot experiment in which RTs for the VO3 condition often included 'vacillating time' between "da" and "ga" after the participants were confident that they were non-labial. We assumed that RTs for VO2 represent time for visual processing which was adequate to cause the McGurk effect (labial vs. non-labial categorization). Therefore, in terms of the visual priming hypothesis, RT differences between AO and VO2 were of our main interest. The VO2 condition was always given just before the VO3 condition. Six practice trials were given before each of the VO2 and VO3 conditions.

Statistical analysis

Group-related effects were mainly examined here as in Experiment 1. ANOVAs for response accuracy were conducted with factors of age group (younger, older) and SNR (low, middle, high) for auditory-related conditions (AO, visual influence calculated by arcAVc − arcAVi), and with factors of task (VO2, VO3) and age group for the VO condition. Similar ANOVAs were done for RT in the AO, AV, and VO conditions as well as for unisensory RT differences (AO − VO). For RTs, the main effect of SNR was also examined for the AO, AVc, and AVi conditions. Response accuracy was also transformed by use of the arcsine function as in Experiment 1. Raw RT data were transformed logarithmically ($\log_{10}$). When significant interaction effects were obtained, *post-hoc* analyses were performed. Planned group comparisons were always conducted for each SNR to examine the main interest of group effect. Greenhouse-Geisser correction was performed when necessary, as in Experiment 1. Lastly, correlation and partial correlation (control variable of hearing threshold) analyses were conducted between RT differences (AO − VO2) and visual influences (AVc − AVi) in all of SNRs for both groups to examine how delayed AO processing contributes to the McGurk effect size.

RESULTS

Percent correct responses

Response accuracy rates for the AO and AV conditions are shown in **Figure 3A** for the younger group and **Figure 3B** for the older group. **Table 2** shows mean response accuracy and statistical results in group comparisons. The response accuracy for the VO condition is also shown in **Figure 3C**. As described below, the older group was not significantly different from the younger in either the auditory (AO) or visual (VO) unisensory conditions, while yielding a larger visual influence in response accuracy (difference between AVc and AVi).

In the AO condition, the main effect of age group or the age group × SNR interaction were not significant [age

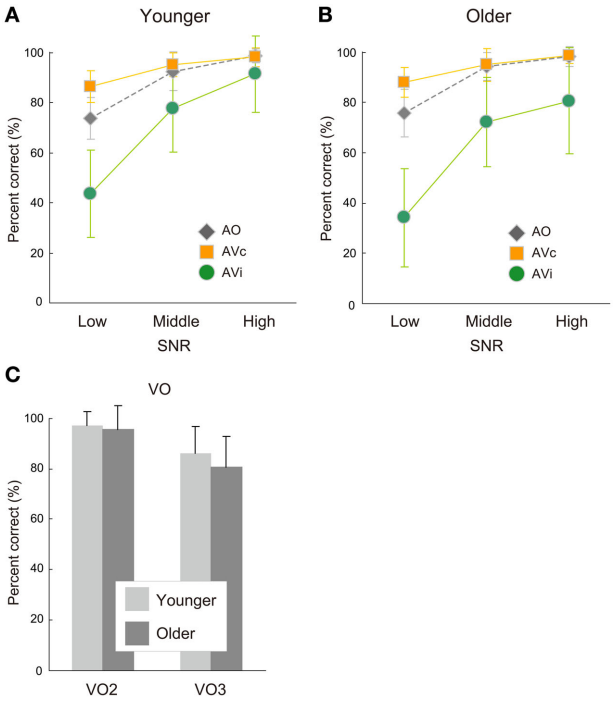


FIGURE 3 | Response accuracy of auditory-only (AO) and auditory-visual (AV) conditions in the (A) younger and (B) older groups and of (C) the visual-only condition with two-alternative (VO2) and three-alternative (VO3) choice in both groups. Response accuracy scores in AV and AO conditions are plotted for the low, middle, and high signal-to-noise ratio (SNR) conditions. AV conditions consisted of two conditions: AV congruent (AVc) and incongruent (AVi) conditions.

Table 2 | Mean response accuracy (%) in Experiment 2 for younger and older groups.

SNR	Younger ($n = 27$)						Older ($n = 17$)						Group difference (p-value)
	AO	AVc	AVi	AVc − AVi	VO2	VO3	AO	AVc	AVi	AVc − AVi	VO2	VO3	AVc − AVi
Low	74	86	43	43	98	91	76	88	34	54	97	82	0.077†
Middle	92	95	78	17			94	95	72	23			0.321
High	99	98	91	7			98	99	81	18			0.018*

SNR, signal-to-noise ratio; AO, auditory-only; AVc, auditory-visual congruent; AVi, auditory-visual incongruent; VO2, visual-only two-alternative choice; VO3, visual-only three alternative choice; $^{\dagger}p < 0.1$; $^{*}p < 0.05$.

group: $F_{(1,\ 42)} = 0.583, p = 0.449, \eta^2 = 0.002$; age group × SNR: $F_{(2,\ 84)} = 0.284,\ p = 0.754,\ \eta^2 = 0.001]$ (**Figures 3A,B**). This indicates that the intelligibility of the auditory stimuli became equivalent for the two groups by successfully manipulating the SNRs.

The VO performances were also similar between the two age groups (**Figure 3C**). Neither the main effect of age group nor the age group × task (VO2, VO3) interaction was significant [age group: $F_{(1,\ 42)} = 2.013, p = 0.163, \eta^2 = 0.001$; age group × task: $F_{(1,\ 42)} = 0.944, p = 0.337, \eta^2 = 0.0003$]. Thus, the two age groups did not statistically differ in terms of lipreading performance. The VO2 task was easier than the VO3 task for both groups [task: $F_{(1,\ 42)} = 83.956, p < 0.0001, \eta^2 = 0.030$].

In contrast, a significant main effect of age group appeared for the visual influence score in AV speech perception [age group: $F_{(1,\ 42)} = 4.990, p = 0.031, \eta^2 = 0.054$; age group × SNR: $F_{(2,\ 84)} = 0.823,\ p = 0.443,\ \eta^2 = 0.006]$ (**Figure 4**). Planned comparisons showed that the older group was more strongly affected by visual information (larger McGurk effect) than the younger, in particular, in the high SNR condition [Bonferroni: low, $p = 0.077$; middle: $p = 0.321$; high: $p = 0.018$].

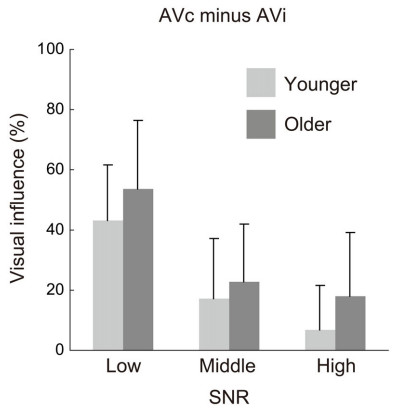


FIGURE 4 | Visual influence scores for the low, middle, and high signal-to-noise ratio (SNR) conditions in the younger and older groups. The scores were calculated by subtraction of percent correct for the auditory-visual incongruent (AVi) condition from the percent correct for the AV-congruent (AVc) condition.

Response time

Mean RTs for each condition are shown in **Figure 5A** for the younger group and **Figure 5B** for the older group. **Table 3** summarizes mean RTs and statistical results in group comparisons. For both age groups, RTs were generally longer for the AVi condition compared with the AVc and AO conditions, replicating the previous results (Sekiyama and Burnham, 2008). The older group showed longer RTs in all conditions except for the VO condition compared with the younger group. Lowering the SNR in audio-related conditions generally tended to lengthen RTs.

In the AO condition, ANOVA found significant main effects of age group $[F_{(1,\ 42)} = 14.800,\ p = 0.0004,\ \eta^2 = 0.216]$ and SNR $[F_{(2,\ 84)} = 16.480;\ p < 0.0001,\ \eta^2 = 0.044]$, while the age group × SNR interaction was not significant $[F_{(2,\ 84)} = 1.942,\ p = 0.150,\ \eta^2 = 0.005]$. Planned group comparisons for each SNR also showed that the older group was generally slower than the younger group (Bonferroni: all of $p < 0.02$), indicating delayed auditory speech perception in older people. Higher SNR conditions tended to be faster than lower SNR conditions across groups [Bonferroni: low vs. middle, $p = 0.051$; low vs. high: $p < 0.0001$; middle vs. high: $p = 0.010$].

As in the AO condition, the older group was slower than the younger group in the AVc condition: ANOVA showed significant

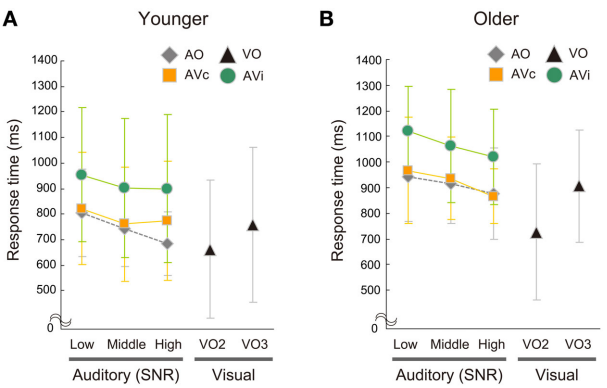


FIGURE 5 | Response times of the auditory-only (AO), auditory-visual (AV), and two-alternative (VO2) and three-alternative (VO3) visual-only conditions in the (A) younger and (B) older groups. AV conditions included two conditions: AV congruent (AVc) and incongruent (AVi) conditions. RTs in the AO and AV conditions are plotted for the low, middle, and high signal-to-noise ratio (SNR) conditions.

main effects of age group $[F_{(1,\ 42)} = 7.129,\ p = 0.011,\ \eta^2 = 0.021]$ and SNR $[F_{(2,\ 84)} = 7.787;\ p = 0.0008,\ \eta^2 = 0.004]$. The age group × SNR interaction was not significant $[F_{(2,\ 84)} = 2.103,\ p = 0.129,\ \eta^2 = 0.001]$. Planned comparisons indicated

Table 3 | Mean response time (ms) in Experiment 2 for younger and older groups.

SNR	Younger ($n = 27$)							Older ($n = 17$)						
	AO	AVc	AVi	VO2	VO3	AO − VO2	AO − VO3	AO	AVc	AVi	VO2	VO3	AO − VO2	AO − VO3
Low	805	822	954	663	759	142	46	942	967	1121	727	906	215	36
Middle	742	761	902			79	−17	918	936	1064			191	12
High	684	773	899			21	−75	877	867	1020			150	−29

	Group difference (p-value)				
SNR	AO	AVc	AVi	AO − VO2	AO − VO3
Low	0.051†	0.019*	0.014*	0.317	0.883
Middle	<0.0001***	0.003**	0.024*	0.165	0.732
High	0.010**	0.067†	0.074†	0.046*	0.522

*SNR, signal-to-noise ratio; AO, auditory-only; AVc, auditory-visual congruent; AVi, auditory-visual incongruent; VO2, visual-only two-alternative choice; VO3, visual-only three alternative choice; †$p < 0.1$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.0001$.*

the older group was significantly or almost significantly slower than the younger group in each SNR condition [Bonferroni: low, $p = 0.019$; middle: $p = 0.003$; high: $p = 0.067$]. The low SNR condition took longer than the middle and high SNR conditions across groups [Bonferroni: low vs. middle, $p = 0.031$; low vs. high: $p = 0.003$; middle vs. high: $p = 0.445$].

In the AVi condition, the older group was also slower than the younger group. Significant main effects of age group [$F_{(1, 42)} = 6.082$, $p = 0.018$, $\eta^2 = 0.105$] and SNR [$F_{(2, 84)} = 5.011$, $p = 0.0124$, $\eta^2 = 0.018$] were found. The age group × SNR interaction was not significant [$F_{(2, 84)} = 0.329$, $p = 0.721$, $\eta^2 = 0.001$, $\varepsilon = 0.861$]. Planned comparisons confirmed that the older group was significantly or almost significantly slower than the younger group in each SNR condition [Bonferroni: low, $p = 0.014$; middle: $p = 0.024$; high: $p = 0.074$]. The low SNR condition took longer than the high SNR condition across groups [Bonferroni: low vs. middle, $p = 0.132$; low vs. high: $p = 0.028$; middle vs. high: $p = 0.787$].

In contrast to the auditory-related conditions, RTs for the VO condition did not show significant group-related differences in either the main effect of age group or the age group × task interaction [age group: $F_{(1, 42)} = 2.394$, $p = 0.129$, $\eta^2 = 0.010$; age group × task: $F_{(1, 42)} = 1.805$, $p = 0.186$, $\eta^2 = 0.001$]. This suggests that the two age groups did not differ in their speed of visual syllable categorization. RT data also showed that the VO2 task was easier than VO3 task for both groups [task: $F_{(1, 42)} = 21.484$, $p < 0.0001$, $\eta^2 = 0.016$].

The differences in RTs between the AO and VO conditions were compared between age groups. Although the main effect of group and the interaction effect of group × SNR were not significant in the ANOVA, the group comparison for the high SNR condition was especially of our interest, because the group difference in the McGurk effect was significantly observed in the high SNR condition in response accuracy. The planned unpaired t-test showed that the temporal difference was larger in the older group than the younger group for the AO relative to VO2 condition in the high SNR condition [$t_{(42)} = 2.055$, $p = 0.046$, Cohen's $d = 0.65$] (**Figure 6**). Such group differences did not reach significance in the middle and low SNR conditions [middle: $t_{(42)} = 1.415$, $p = 0.165$, Cohen's $d = 0.45$; low: $t_{(42)} = 1.012$, $p = 0.317$, Cohen's $d = 0.32$]. There were no significant group differences in AO − VO3 [high: $t_{(42)} = 0.645$, $p = 0.522$, Cohen's $d = 0.20$; middle: $t_{(42)} = 0.344$, $p = 0.732$, Cohen's $d = 0.11$; low: $t_{(42)} = 0.149$, $p = 0.883$, Cohen's $d = 0.05$].

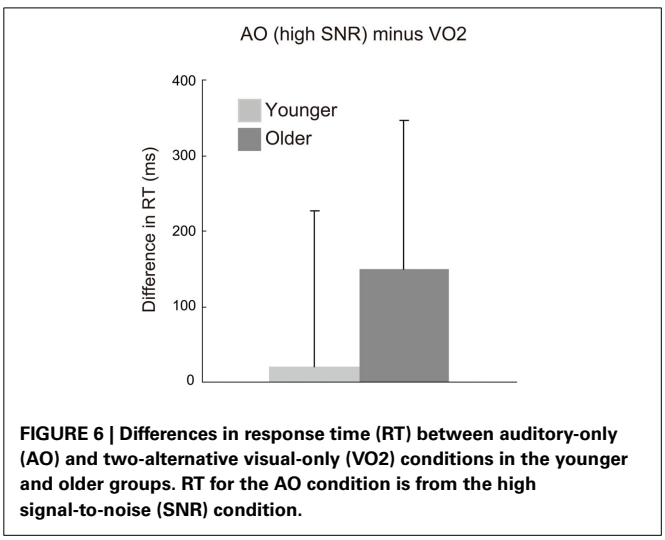


FIGURE 6 | Differences in response time (RT) between auditory-only (AO) and two-alternative visual-only (VO2) conditions in the younger and older groups. RT for the AO condition is from the high signal-to-noise (SNR) condition.

Finally, continuous correlation analyses were conducted between RT differences (AO − VO2) and percent visual influence scores (AVc − AVi) for both the younger and older groups, using all of three SNR data. Although the younger group did not show a significant correlation [$r = 0.127$, $p = 0.258$; $n = 81$] (**Figure 7A**), the older group yielded a significant positive correlation [$r = 0.337$, $p = 0.016$; $n = 51$] (**Figure 7B**). Such significant correlation relationship for the older group remained significant under the control of hearing thresholds [older ($n = 48$): $\rho_{XY.Z} = 0.374$, $p = 0.008$; younger ($n = 78$): $\rho_{XY.Z} = 0.128$, $p = 0.259$]. As inferred from this, there was no significant correlation between the RT differences (AO − VO2) and hearing thresholds in the older group ($r = -0.094$, $p = 0.512$). These results indicate that

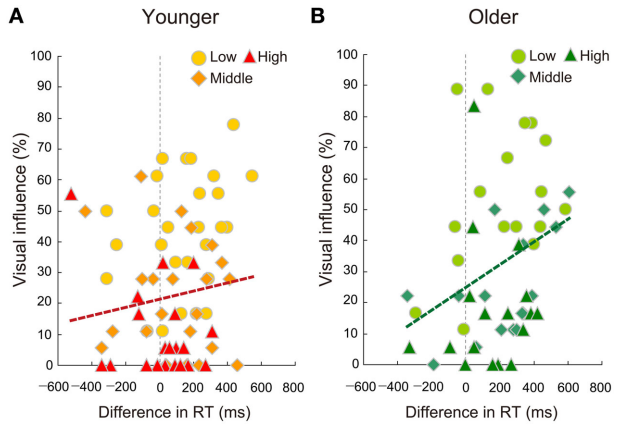


FIGURE 7 | Relationships between difference in RT (auditory-only minus two-alternative visual-only conditions) and visual influence in percent correct (auditory-visual congruent minus incongruent conditions) using all data from the three signal-to-noise ratios (Low, Middle, High) in the (A) younger and (B) older groups. The older group, but not the younger group, showed a significant correlation.

more delayed AO perception (being positive in difference in RT) was related with larger McGurk effects in the older group.

In summary, in discrete analyses, the AO − VO2 RT difference was significantly larger for the older group in the high SNR condition, and this coincided with the result for the visual influence score for which planned comparisons showed a significant age difference in the high SNR condition. In continuous analyses across all of the SNR conditions, the older group showed a significant correlation between the size of the McGurk effect and the unisensory RT difference (AO − VO2), indicating that the larger McGurk effect is associated with more delayed AO perception. These results support the visual priming hypothesis.

DISCUSSION

Experiment 2 compared older and younger participants not only in terms of response accuracy, but also RT; therefore, we calibrated the SNR of auditory stimuli so that the auditory intelligibility was equivalent for both age groups. With a difference of 4 dB of SNR between the two age groups, the older and the younger were tested in low (1 or −3 dB), middle (11 or 7 dB), and high (21 or 17 dB) SNRs in the AV and AO conditions. The results showed that the McGurk effect was still stronger for older than for younger adults under equivalent auditory intelligibility. The two age groups also showed equivalent accuracy of VO performance. Because the two age groups were equivalently accurate in both AO and VO performances, the age difference in the McGurk effect needs to be explained by a factor other than unisensory accuracy.

Response times revealed that the delay due to aging was large in conditions that included auditory stimuli (AO, AVi, and AVc), whereas there was no such delay in lipreading; this was especially so for labial–non-labial categorization (VO2). Because we assumed that RTs for VO2 represent time for visual processing which was adequate to cause the McGurk effect (labial–non-labial categorization), we focused on the visual precedence time in the binary lipreading condition (AO − VO2 in RT). The visual precedence time (AO − VO2) was significantly larger for the older group than the younger group in the high SNR condition, but not in the middle and low SNR conditions. This was in accordance with the fact that the aging-related increase in the visual influence on accuracy tended to be more pronounced for the high SNR condition. These results suggest that the older participants' larger visual precedence due to delayed auditory processing (particularly in the high SNR condition) is related to a larger visual influence. The co-occurrence of the larger visual precedence and the larger visual influence in the older group is consistent with our visual priming hypothesis.

Moreover, the within-group correlation analysis across all SNRs found a significant correlation between the size of the McGurk effect and the unisensory RT difference (AO − VO2) in the older adults: The larger McGurk effect was associated with the larger visual precedence, supporting the visual priming hypothesis. Such an association was not found in the younger participants, thus the association in the older participants seems to be based on the aging-related auditory delay.

On the other hand, the visual precedence time for three-alternative lipreading conditions (AO − VO3 in RT) was not significantly different between the two age groups. This may be a general tendency of the elderly who attach importance to accuracy rather than speed when the task is difficult (in VO3).

GENERAL DISCUSSION

This study investigated whether or not older adults with normal hearing and preserved lipreading use more visual speech information than younger adults in auditory-visual speech perception. Particularly, we intended to examine our visual priming hypothesis that emphasizes the amount of temporal precedence of VO speech processing relative to AO processing as a cause of the aging-related increase in visual influence.

Previous studies on aging-related differences in auditory-visual speech perception presented auditory stimuli to older and younger adults either under the same SNRs (Thompson, 1995; Behne et al., 2007; Setti et al., 2013) or calibrated SNRs (Cienkowski and Carney, 2002; Sommers et al., 2005), and only the same-SNR settings found significant aging-related differences. Among the above studies, only some studies conducted screening of the participants based on hearing thresholds (Cienkowski and Carney, 2002; Sommers et al., 2005; Setti et al., 2013). Concerning the age range of the participants, control was not so strict in most of these studies. In fact, studies including older adults over 70 years have often revealed poorer lipreading in older adults, which would make it complicated to assess aging-related changes in AV integration. Our strategies were (1) to use both the same SNRs and calibrated SNRs, (2) to exclude participants with clinically declined hearing, and (3) to minimize the aging-related decline in lipreading by setting an age range of older adults between 60 and 65 years.

We found that the visual influence was greater in the older adults compared with the young adults not only in the same SNRs, but also in the calibrated SNRs. Based on the effect size of the main effect of age group ($\eta^2 = 0.188$ in Experiment 1; $\eta^2 = 0.105$ in Experiment 2), the aging-related difference in the visual influence were larger under the same SNRs than the calibrated

SNRs. This is reasonable because the same-SNR setting did not correct the aging-related poorer AO performance for the older adults, so it would have led to a greater visual influence on them as predicted from optimal integration models (Massaro, 1987, 1998; Braida, 1991; Grant et al., 1998; Schwartz, 2010).

The novel finding of the present study is that the aging-related increase in the visual influence was significant even under the calibrated SNRs. Importantly, the calibration was successful as confirmed by non-significant age group differences for unisensory AO accuracy. Therefore, for the first time, the aging-related increase in visual influence was revealed after controlling for the hearing decline of older adults. In the accuracy data in Experiment 2, there were no age group differences in unisensory performance not only in the AO, but also in the VO conditions. Nevertheless, the multisensory AV integration differed between the two age groups. Therefore, the differential AV integration between the two age groups must be attributable to some factors other than unisensory accuracy: this is a starting point to examine our visual priming hypothesis, which was supported by the present RT results.

The RT difference between older and younger adults was constant in audio-related conditions (AO and AV), while no such delay in RTs for older adults relative to younger adults was observed in the VO condition. Of importance, this aging-related auditory delay could be persistent when the visual labial–nonlabial decision (VO2) was not delayed. Thus, the older group's larger RTs in the AO condition were not attributable to general response slowing, but to the modality-specific delay in auditory processing. Consequently, the visual precedence time (AO − VO2) was significantly longer in the older than the younger adults in the high SNR condition. In accordance with this, the aging-related increase in visual influence tended to be more pronounced in the high SNR condition, yielding a larger McGurk effect in the older adults. Moreover, the correlation analyses within the older group across SNRs indicated that more delayed RT is associated with the larger McGurk effect. Therefore, the visual priming hypothesis was supported in two aspects: One is the group differences in the high SNR condition, and the other is the correlation within the older group.

The delayed auditory processing of older adults has also been found in studies using ERPs both for speech (Tremblay and Ross, 2007) and non-speech (Schroeder et al., 1995). Furthermore, ERPs for AV congruent stimuli revealed that the temporal facilitation of speech processing by visual speech is greater for normal hearing older adults compared with younger adults (Winneke and Phillips, 2011). Such a temporal, visual facilitation is thought to be due to anticipation provided by visual lipread information that starts a few hundred milliseconds earlier than the onset of auditory energy in natural speech articulation (Van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009). The temporal facilitation in ERPs was also observed for non-speech events where the anticipatory visual motion precedes the sound, for example, in hand clapping, but not in events where visual motion and sound start at the same time, for example, paper tearing (Stekelenburg and Vroomen, 2007). Thus, anticipatory visual motion may predict when a sound will occur (Stekelenburg and Vroomen, 2007) and what phonemes are candidates (Van Wassenhove et al., 2005). The visual precedence time in RT in the present older adults could be a measure of temporal information about how much in advance the visual anticipation is generated relative to the auditory perception. The present results suggest that visual anticipation may function well to influence auditory processing, when visual precedence time is at least about 100 ms, as observed in the RT difference between AO and VO2 in the older adults.

Concerning SNRs and aging-related performance differences, the relationship between AO accuracy and the visual influence (AVc − AVi) was not always simple. In experiment 1, significant group differences in AO accuracy were found in two SNRs, while group differences in the visual influence score were significant at all SNR levels. This seems in accordance with the fact that the effect of lipreading on AV accuracy is not additive to AO accuracy, but in a multiplicative way (e.g., Braida, 1991), thus, small or non-significant differences in AO conditions could turn into large differences in AV conditions (Sumby and Pollack, 1954). In Experiment 2, we used calibrated SNRs to eliminate group differences in AO accuracy, thus it is naturally expected that group differences in the visual influence would be observed in more limited way compared with Experiment 1. In fact, a significant group difference in visual influence score was found only at the high SNR.

It was unexpected that the group difference was more prominent at the high SNR than the middle and low SNRs. Why was this? It may have been due to the relativity in RTs between the AO and VO conditions. Although the older group showed a constant AO delay relative to the younger group at each SNR, the RTs became longer as the SNR became lower for both groups. As a result, the visual precedence time (AO − VO2), which was almost zero for the younger adults at the high SNR, reached a substantial amount at the middle and low SNRs for the younger, as well as for the older, adults (**Figure 5**). This caused a substantial degree of visual influence on both groups in the middle and low SNRs (**Figure 4**), which may have resulted in reduced age group differences.

On the other hand, there may be a case in which the visual priming hypothesis does not hold. In the context of non-speech processing, a previous study demonstrated that multisensory facilitation on RT of simple detection relative to unisensory detection was greater for older adults than young adults (Peiffer et al., 2007). They used lights and white noise as stimuli, and a multisensory condition was presented to them at the same time. An aging-related increase in multisensory facilitation was still found even when unisensory detection was equally fast for both age groups. The time course in which visual and auditory streams are integrated may be different depending on stimuli (dynamic visual motion vs. static light, and anticipatory vs. abrupt visual cues) and task (categorization vs. detection).

Recently, individual differences in the McGurk effect among young perceivers were studied in terms of the "temporal binding window" (Stevenson et al., 2012). These authors found that persons who are more sensitive to beep-flash asynchrony (thus with smaller temporal binding window) are more susceptible to the McGurk effect. This suggests that mechanisms for detecting auditory-visual simultaneity are also relevant to some extent for integration of auditory and visual speech information. Could

older adults with delayed auditory processing have any drawbacks to auditory-visual simultaneity detection? One possibility is that the delay of auditory relative to visual processing may be perceptually canceled as the older adults adapt to the aging-related delay and recalibration takes place as found for experimental lags in young adults (Fujisaki et al., 2004). If so, the temporal binding window itself may not be a source of aging-related differences in the McGurk effect. However, the extent to which the temporal binding window accounts for individual differences in the McGurk effect may differ between age groups. In the present study, the visual precedence (that is, auditory delay) was associated with the size of the McGurk effect only in the older adults. Therefore, the young adults' individual differences in the McGurk effect should be accounted for by the other factors, such as the temporal binding window, whereas those of the older adults are possibly accounted for by both the auditory delay and temporal binding window.

Finally, we should mention the inconsistency between the present findings of a larger McGurk effect in the older group and the previous findings (Cienkowski and Carney, 2002; Sommers et al., 2005). A few factors may have contributed to the inconsistency. One is the age range of the participants: we excluded those over 66 years to minimize lipreading decline (Shoop and Binnie, 1979). Another critical difference may be the range of SNRs: we used a wider range of SNRs including much milder SNRs compared with the previous studies. These factors may have partially contributed to the inconsistency between the present and previous studies.

In conclusion, this study demonstrated that native Japanese speaking older adults used more visual speech information than their younger counterparts, and were more susceptible to the McGurk effect when tested with stimuli containing equivalently intelligible auditory speech. From the RT data, the enhanced visual influence on the older adults was likely associated with an aging-related delay in auditory processing. The delay was observed despite the equalized AO accuracy between the two age groups, presumably representing aging-related changes in higher order neural processes that are hard to observe by hearing thresholds alone (Pichora-Fuller and MacDonald, 2009). Time-related measures such as RTs and ERPs are important to assess older adults' auditory perception. In this study, there was no correlation between hearing thresholds and delay in auditory RT, indicating that the two factors are dissociable. Thus, among the older adults with normal hearing, it may be that the delay in cortical auditory processing, rather than peripheral sensory sensitivity, is more critical for the greater visual influence. Furthermore, it was previously shown that the RT difference between auditory and visual speech perception was larger for young native English speakers than for young Japanese speakers (Sekiyama and Burnham, 2008). It will be of interest to clarify in the future whether or not the procedure used in the present study can reveal an aging-related increase in visual precedence in English speaking populations as in Japanese.

ACKNOWLEDGMENTS

This work was supported by Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (No. 18530563, No. 21243040, and No. 25245068 to Kaoru Sekiyama). The authors are grateful for the support provided by Yuji Takahashi, Ayaka Inoda, Hideki Kobayashi, and Kie Takigawa.

REFERENCES

Altieri, N., Pisoni, D. B., and Townsend, J. T. (2011). Some behavioral and neurobiological constraints on theories of auditory-visual speech integration: a review and suggestions for new directions. *Seeing Perceiving* 24, 513–539. doi: 10.1163/187847611X595864

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Bailly, G., Perrier, P., and Vatikiotis-Bateson, E. (eds.). (2012). *Auditory-visual Speech Processing*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511843891

Behne, D., Wang, Y., Alm, M., Arntsen, I., Eg, R., and Valso, A. (2007). "Changes in auditory-visual speech perception during adulthood," in *Proceedings of the International Conference on Auditory-Visual Speech Processing 2007*, eds J. Vroomen, M. Swerts, and E. Krahmer (Hilvarenbeek).

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Q. J. Exp. Psychol. A* 43, 647–677. doi: 10.1080/14640749108400991

Campbell, R., Dodd, B., and Burnham, D. (eds.). (1998). *Hearing by Eye II*. Hove: Psychology Press.

CHABA, Committee on Hearing, Bioacoustics, and Biomechanics (Tobias, J. V. et al.). (1988). Speech understanding and aging. *J. Acoust. Soc. Am.* 83, 859–895. doi: 10.1121/1.395965

Chen, Y., and Hazan, V. (2009). Developmental factors and the non-native speaker effect in auditory-visual speech perception. *J. Acoust. Soc. Am.* 126, 858–865. doi: 10.1121/1.3158823

Cienkowski, K. M., and Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear Hear.* 23, 439–449. doi: 10.1097/00003446-200210000-00006

Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778. doi: 10.1038/nn1268

Glorig, A., and Nixon, J. (1962). Hearing loss as a function of age. *Laryngoscope* 72, 1596–1610. doi: 10.1288/00005537-196211000-00006

Grant, K. W., and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208. doi: 10.1121/1.1288668

Grant, K. W., Walden, B. E., and Seits, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 103, 2677–2690. doi: 10.1121/1.422788

Howell, D. C. (1997). *Statistical Methods for Psychology.* 4th Edn. Belmont, CA: Wadsworth Publishing.

Kuhl, P. K., Tsuzaki, M., Tohkura, Y., and Meltzoff, A. N. (1994). "Human processing of auditory-visual information in speech perception: potential for multimodal human-machine interfaces," in *Proceedings of the International Conference of Spoken Language Processing,* ed The Acoustical Society of Japan (Tokyo: The Acoustical Society of Japan), 539–542.

Luce, R. D. (1986). *Response Times*. New York, NY: Oxford University Press.

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: The MIT Press.

Massaro, D. W., Thompson, L. A., Barron, B., and Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *J. Exp. Child Psychol.* 41, 93–113. doi: 10.1016/0022-0965(86)90053-6

Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., and Heredia, R. (1993). Bimodal speech perception: an examination across languages. *J. Phon.* 21, 445–478.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Peiffer, A. M., Mozolic, J. L., Hugenschmidt, C. E., and Laurienti, P. J. (2007). Age-related multisensory enhancement in a simple audiovisual detection task. *Neuroreport* 18, 1077–1081. doi: 10.1097/WNR.0b013e3281e72ae7

Pichora-Fuller, M. K., and MacDonald, E. (2009). "Sensory aging: hearing," in *Handbook of the Neuroscience of Aging,* eds P. R. Hof and C. V. Mobbs (London: Academic Press), 193–198.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

Rouger, J., Fraysse, B., Deguine, O., and Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Res.* 1188, 87–99. doi: 10.1016/j.brainres.2007.10.049

Schorr, E. A., Fox, N. A., van Wassenhove, V., and Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18748–18750. doi: 10.1073/pnas.0508862102

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* 12, 106–113. doi: 10.1016/j.tics.2008.01.002

Schroeder, M. M., Lipton, R. B., Ritter, W., Giesser, B. S., and Vaughan, H. G. Jr., (1995). Event-related potential correlates of early processing in normal aging. *Int. J. Neurosci.* 80, 371–382. doi: 10.3109/00207459508986110

Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that auditory-visual fusion in speech perception is subject-dependent. *J. Acoust. Soc. Am.* 127, 1584–1594. doi: 10.1121/1.3293001

Schwartz, J. L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early auditory-visual interactions in speech identification. *Cognition* 93, 69–78. doi: 10.1016/j.cognition.2004.01.006

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *J. Acoust. Soc. Jpn.* 15, 143–158. doi: 10.1250/ast.15.143

Sekiyama, K., and Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Dev. Sci.* 11, 306–320. doi: 10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neurosci. Res.* 47, 277–287. doi: 10.1016/S0168-0102(03)00214-1

Sekiyama, K., and Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am.* 90, 1797–1805. doi: 10.1121/1.401660

Sekiyama, K., and Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *J. Phon.* 21, 427–444.

Setti, A., Burke, K. E., Kenny, R., and Newell, F. N. (2013). Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes. *Front. Psychol.* 3:4. doi:10.3389/fpsyg.2013.00575

Shoop, C., and Binnie, C. A. (1979). The effects of age upon the visual perception of speech. *Scand. Audiol.* 8, 3–8. doi: 10.3109/01050397909076295

Sommers, M. S., Tye-Murray, N., and Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear.* 26, 263–275. doi: 10.1097/00003446-200506000-00003

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid auditory-visual events. *J. Cogn. Neurosci.* 19, 1964–1973. doi: 10.1162/jocn.2007.19.12.1964

Stevenson, R. A., Zemtsov, R. K., and Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1517–1529. doi: 10.1037/a0027339

Stork, D. G., and Hennecke, M. E. (eds.). (1996). *Lipreading by Humans and Machines: Models, Systems, and Applications*. Berlin: Springer-Verlag. doi: 10.1007/978-3-662-13015-5

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Thompson, L. A. (1995). Encoding and memory for visible speech and gestures: a comparison between young and older adults. *Psychol. Aging* 10, 215–228. doi: 10.1037/0882-7974.10.2.215

Thompson, L. A., and Malloy, D. M. (2004). Attention resources and visible speech encoding in older and younger adults. *Exp. Aging Res.* 30, 241–252. doi: 10.1080/03610730490447877

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., and Theoret, H. (2007). Speech and non-speech audio-visual illusions: a developmental study. *PLoS ONE* 2:e742. doi: 10.1371/journal.pone.0000742

Tremblay, K., and Ross, B. (2007). Effects of age and age-related hearing loss on the brain. *J. Commun. Disord.* 40, 305–312. doi: 10.1016/j.jcomdis.2007.03.008

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

Walden, B. E., Busacco, D. A., and Montgomery, A. A. (1993). Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons. *J. Speech Hear. Res.* 36, 431–436.

WHO definition of hearing impairment. Available online at: http://www.who.int/pbd/deafness/hearing_impairment_grades/en/index.html

Winneke, A. H., and Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in auditory-visual speech perception. *Psychol. Aging* 26, 427–438. doi: 10.1037/a0021683

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 February 2014; paper pending published: 12 March 2014; accepted: 28 March 2014; published online: 14 April 2014.

Citation: Sekiyama K, Soshi T and Sakamoto S (2014) Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. Front. Psychol. ***5****:323. doi: 10.3389/fpsyg.2014.00323*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 21 May 2014
doi: 10.3389/fpsyg.2014.00379

The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders

Ryan A. Stevenson[1]*, Magali Segers[2], Susanne Ferber[1], Morgan D. Barense[1,3] and Mark T. Wallace[4,5,6,7,8]

[1] *Department of Psychology, University of Toronto, Toronto, ON, Canada*
[2] *Department of Psychology, York University, Toronto, ON, Canada*
[3] *Rotman Research Institute, Toronto, ON, Canada*
[4] *Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN, USA*
[5] *Vanderbilt University Medical Center, Vanderbilt Brain Institute, Nashville, TN, USA*
[6] *Vanderbilt Kennedy Center, Nashville, TN, USA*
[7] *Department of Psychology, Vanderbilt University, Nashville, TN, USA*
[8] *Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, USA*
**Correspondence: ryan.andrew.stevenson@gmail.com*

Edited by:
Jean-Luc Schwartz, Centre National de la Recherche Scientifique, France

Reviewed by:
Julia Irwin, Haskins Laboratories, USA

Keywords: autism spectrum disorders (ASD), autism, multisensory integration, audiovisual, audiovisual processing, developmental disabilities, sensory perception, speech perception

Speech perception is an inherently multisensory process. When having a face-to-face conversation, a listener not only hears what a speaker is saying, but also sees the articulatory gestures that accompany those sounds. Speech signals in visual and auditory modalities provide complementary information to the listener (Kavanagh and Mattingly, 1974), and when both are perceived in unison, behavioral gains in in speech perception are observed (Sumby and Pollack, 1954). Notably, this benefit is accentuated when speech is perceived in a noisy environment (Sumby and Pollack, 1954). To achieve a behavioral gain from multisensory processing of speech, however, the auditory and visual signals must be perceptually bound into a single, unified percept. The most commonly cited effect that demonstrates perceptual binding in audiovisual speech perception is the McGurk effect (McGurk and MacDonald, 1976), where a listener *hears* a speaker utter the syllable "ba," and *sees* the speaker utter the syllable "ga." When these two speech signals are perceptually bound, the listener perceives the speaker as having said "da" or "tha," syllables that are not contained in either of the unisensory signals, resulting in a perceptual binding, or integration, of the speech signals (Calvert and Thesen, 2004).

The ability to perceptually bind sensory information is notably impaired in a number of clinical populations, including those with autism spectrum disorders (ASD). ASD describes a cluster of highly prevalent developmental disabilities historically characterized by deficits in three functional domains: language and communication, social reciprocity, and the presence of restricted interests/repetitive behaviors (APA, 2000). Since its initial description, alterations in sensory processing have been described in this population (Kanner, 1943), yet these deficits were acknowledged only in the most recent edition of the DSM (APA, 2013). Impairments in multisensory perceptual binding may be particularly relevant in ASD, given that hallmark features of the disorder include difficulties in speech, communication, and social interactions. Successful speech communication is heavily reliant on binding across sensory modalities, and as such, impaired binding in individuals with ASD likely contributes to these core deficits.

Impairments in perceptual binding have not gone unstudied in ASD. In fact, one of the leading theories describing ASD, Weak Central Coherence, describes ASD as a cognitive style in which focus is selectively attuned to individual components of information to the exclusion of perceiving the larger whole; in short, losing the proverbial forest for the trees (Frith and Happé, 1994; Happé, 1999, 2005; Happé and Frith, 2006). Evidence for this has been found across a wide range of tasks. For example, individuals with ASD benefit less than individuals without ASD from context when interpreting a sentence or story (Happé, 1994; Jolliffe and Baron-Cohen, 1999), but are more accurate than individuals without ASD when focusing on explicit local details of a passage (Noens and Berckelaer-Onnes, 2005).

In the realm of sensory perception, binding deficits in ASD have been studied most extensively in the visual modality. Here too, individuals with ASD have been shown to have a strong local bias at the expense of global processing (Behrmann et al., 2006). A clear example of this is observed in response to hierarchical letters (large letters composed of smaller letters; Navon, 1977). When performing a task reliant upon the identify the gestalt of the image (the large letter) relative to the individual units (small component letters), individuals with ASD show impaired performance (Behrmann et al., 2006).

The ability of individuals with ASD to bind *across* sensory modalities has been studied to a much lesser extent, but those studies that have been conducted commonly find deficits in multisensory perceptual binding, particularly with speech signals. The majority of the research suggests that individuals with ASD perceive the McGurk illusion less often than their

peers without ASD (de Gelder et al., 1991; Williams et al., 2004; Mongillo et al., 2008; Irwin et al., 2011; Bebko et al., 2014; Stevenson et al., 2014, in press; but see Iarocci and McDonald, 2006; Woynaroski et al., 2013), often relying instead on the auditory modality to the exclusion of the visual information (Mongillo et al., 2008; Stevenson et al., 2014, in press). While individuals with ASD may be able to perceptually bind information under optimal conditions, these results imply that individuals with ASD show reduced efficiency when binding speech information across auditory and visual modalities, particularly in noisy, real-world contexts (Foxe et al., 2013). As a consequence, signals are perceived in isolation, or as fragmented units rather than as a meaningful whole. Thus, the efficiency gained from processing multiple sensory signals as a single percept, for example the visual sensory inputs associated with a speaker integrated with the auditory sensory inputs associated with a speaker (Stevenson et al., 2010, 2011), would be lost, resulting in more inefficient sensory processing overall.

Given the findings that individuals with ASD show reduced perceptual binding of audiovisual speech signals, it has been hypothesized that individuals with ASD would not exhibit the behavioral gains observed with the perception of multisensory signals. The few studies to date that have investigated multisensory perception of audiovisual speech have shown that children with ASD do in fact show less behavioral gain (i.e., less improved perception) with audiovisual speech than do their typically developing peers (Alcántara et al., 2004; Smith and Bennetto, 2007; Irwin et al., 2011; Foxe et al., 2013). This finding is especially salient when speech is embedded in a high degree of background noise (Foxe et al., 2013), the very condition in which (A) typically developing children show a high level of multisensory gain and (B), this multisensory integration would be most beneficial for successful speech communication. The validity of the relationship between multisensory perception and real-world communication has been demonstrated via correlations between the accurate perception of audiovisual speech and communication scores from the Autism Diagnostic Observation Schedule (Lord et al., 2000), the gold standard for diagnostic testing in ASD. Individuals who were better able to accurately perceive audiovisual speech were less impaired in terms of communicative abilities (Woynaroski et al., 2013).

Interestingly, multisensory speech integration is not a static process, but one that continues to mature and fine tune over development (Hillock et al., 2011; Hillock-Dunn and Wallace, 2012). While young children with ASD are clearly delayed in their ability to benefit from multisensory speech perception compared to their typically developing peers, there is evidence that this impairment lessens with maturation (Foxe et al., 2013). Likewise, the first study of the McGurk Effect across development showed a similar pattern, in which young children with ASD perceived the McGurk Effect much less frequently than their peers without ASD, but "caught up" later in development (Taylor et al., 2010; but see Stevenson et al., in press).

A critical question then, is what is the underlying cause of these disruptions in speech perception observed in ASD? One possibility is that individuals with ASD have impaired temporal processing abilities. One neurobiological account of ASD, the temporal binding hypothesis of autism (Brock et al., 2002) proposes just that. In terms of binding across sensory inputs, perceiving the timing of incoming sensory information is paramount to the ability to perceptually bind stimuli across sensory modalities. The temporal synchrony of such inputs is one, if not the most, salient cue that two inputs *should* be bound (Vroomen and Keetels, 2010). Previous research shows a clear pattern that individuals with ASD are significantly impaired in judging the relative timing of auditory and visual speech signals (Bebko et al., 2006; Foss-Feig et al., 2010; Kwakye et al., 2011; de Boer-Schellekens et al., 2013; Woynaroski et al., 2013; Stevenson et al., 2014), and importantly, this research also showed a direct correlation between multisensory temporal acuity and the ability to perceptually bind audiovisual speech signals in individuals with ASD (Stevenson et al., 2014).

These findings, taken in sum, suggest that deficits in binding across auditory and visual modalities in ASD may have a cascading impact on speech perception and social processing, key clinical symptoms defining ASD. In most social communicative interactions, failing to perceive the auditory and visual components of the environment can result in missing critical social cues, not to mention the content of the message being conveyed. Failing to perceive a speaker's message as a single, unified percept, essentially doubles the number of perceived inputs, resulting in an increasingly "noisy" or "intense" world—as is often described in the case of autism (Just et al., 2004; Markram et al., 2007; Rippon et al., 2007; Pouget et al., 2009).

The impact of an inability to perceptually bind across senses on other aspects of cognition has been well characterized in a patient with bilateral parietal hypoperfusion (Hamilton et al., 2006). This patient, AWF, began to perceive what he heard and what he saw as being out of sync. As a result of this atypical multisensory temporal processing, AWF was unable to perceptually bind audiovisual speech, indexed by an inability to perceive the McGurk Effect. Additionally, AWF no longer showed the typical behavioral benefits with he was shown a speakers mouth and articulatory gestures accompanying auditory speech. While the etiology of AWF's impairment is clearly distinct from ASD, the parallels in the perception of audiovisual speech are striking. Furthermore, AWF's describes coping with his asynchronous environment by limiting face-to-face conversations and looking away from the face during in-person conversations, both behaviors commonly seen in ASD. Such a coping strategy may reflect the perceived avoidance of social interactions in ASD, which may relate more to limiting the amount of perceptual noise in the environment. A similar argument has been made for self-stimulation or "stimming" behaviors commonly observed in ASD. It is possible that these repetitive movements provide a predictable and controlled sensory experience in an otherwise chaotic world (Jones et al., 2003).

While the impact that atypical sensory binding appears to have on the core symptoms associated with ASD is supported by research, the issue of how to translate these findings into clinical practice has been largely unexplored (note here that treatments commonly referred to as "sensory integration therapy" do

not in fact focus on binding or integrating information across sensory modalities). Intensive Behavioral Intervention (IBI) is the evidence-based treatment of choice for ASD; however, the degree of gain made by any one child is difficult to predict. While milder autism severity, higher adaptive functioning, and higher cognitive skills are related to better outcomes, there remain unaccounted for factors which may predict which children benefit most from treatment (Flanagan et al., 2012). Given that sensory and multisensory processing are foundational to the higher-level cognitive, communicative, and social functioning that treatments aim to address, knowledge of an individual's ability to process sensory information is a critical and necessary first step to benefit maximally from intensive intervention.

These possible clinical implications are, at this stage, highly speculative. The possible upsides, however, of moving this research from the laboratory into real-world settings are significant. A clear consensus of evidence suggests that individuals with ASD process and integrate sensory information in an atypical manner, and that this is strongly linked to core impairments in communicative and social abilities. A number of research questions must be addressed in order to explore these possibilities. First, longitudinal studies of individuals with ASD need to be conducted to directly assess how speech and communication skills develop in conjunction with sensory processing, specifically binding across sensory modalities and multisensory temporal processing. Second, the mediating or moderating effect that specific sensory-processing phenotypes in ASD have on the efficacy of evidence-based treatments such as IBI is sorely needed (in addition to other variables such as IQ and gender; Wolery and Garfinkle, 2002; Rogers and Vismara, 2008). Finally, research should ultimately go beyond documenting the sensory and multisensory processing abilities of individuals with ASD and in addition, should also reveal how these abilities can be dynamically modulated. Plasticity within the relevant perceptual systems has been amply demonstrated (Fujisaki et al., 2004; Powers et al., 2009; Stevenson et al., 2013; Schlessinger et al., in press), but these findings have been not yet been applied to populations with ASD. Pursuing these and related studies has the potential to not only add to our understanding of ASD, but also, through clinical application, to improve the quality of life of individuals with ASD.

ACKNOWLEDGMENTS

Funding for this work was provided by a Banting Postdoctoral Fellowship administered by the Government of Canada *It's only a matter of time: Neural networks underlying multisensory perceptual binding*, a University of Toronto Department of Psychology Postdoctoral Fellowship Grant, National Institutes of Health F32 DC011993 *Multisensory Integration and Temporal Processing in ASD*, National Institutes of Health R34 DC010927 *Evaluation of Sensory Integration Treatment in ASD*, National Institutes of Health R21 CA1834892 *Multisensory Processing Across Lifespan and Links to Cognition*, a Simons Foundation research grant *Exploring Links Between Multisensory and Cognitive Function in Autism*, a Vanderbilt Institute for Clinical and Translational Research grant VR7263 *Neuroplasticity of Sensory Processing in Autism Spectrum Disorders*, a Vanderbilt Kennedy Center MARI/Hobbs Award, the Vanderbilt Brain Institute, and the Vanderbilt University Kennedy Center.

REFERENCES

Alcántara, J. I., Weisblatt, E. J., Moore, B. C., and Bolton, P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *J. Child Psychol. Psychiatry* 45, 1107–1114. doi: 10.1111/j.1469-7610.2004.t01-1-00303.x

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders-IV-TR.* Washington, DC: APA.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5).* American Psychiatric Association.

Bebko, J. M., Schroeder, J. H., and Weiss J. A. (2014). *The McGurk Effect in Children With Autism and Asperger Syndrome*, Vol. 7, Autism Research. 50–59.

Bebko, J. M., Weiss, J. A., Demark, J. L., and Gomez, P. (2006). Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *J. Child Psychol. Psychiatry* 47, 88–98. doi: 10.1111/j.1469-7610.2005.01443.x

Behrmann, M., Avidan, G., Leonard, G. L., Kimchi, R., Luna, B., Humphreys, K., et al. (2006). Configural processing in autism and its relationship to face processing. *Neuropsychologia* 44, 110–129. doi: 10.1016/j.neuropsychologia.2005.04.002

Brock, J., Brown, C. C., Boucher, J., and Rippon, G. (2002). The temporal binding deficit hypothesis of autism. *Dev. Psychopathol.* 14, 209–224. doi: 10.1017/S0954579402002018

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

de Boer-Schellekens, L., Eussen, M., and Vroomen, J. (2013). Diminished sensitivity of audiovisual temporal order in autism spectrum disorder. *Front. Integr. Neurosci.* 7:8. doi: 10.3389/fnint.2013.00008

de Gelder, B., Vroomen, J., and Van der Heide, L. (1991). Face recognition and lip-reading in autism. *Eur. J. Cogn. Psychol.* 3, 69–86. doi: 10.1080/09541449108406220

Flanagan, H. E., Perry, A., and Freeman, N. L. (2012). Effectiveness of large-scale community-based intensive behavioral Intervention: a waitlist comparison study exploring outcomes and predictors. *Res. Autism Spectr. Disord.* 6, 673–682. doi: 10.1016/j.rasd.2011.09.011

Foss-Feig, J. H., Kwakye, L. D., Cascio, C. J., Burnette, C. P., Kadivar, H., Stone, W. L., et al. (2010). An extended multisensory temporal binding window in autism spectrum disorders. *Exp. Brain Res.* 203, 381–389. doi: 10.1007/s00221-010-2240-4

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H. P., Russo, N. N., Blanco, D., et al. (2013). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cereb. Cortex*. doi: 10.1093/cercor/bht213. [Epub ahead of print].

Frith, U., and Happé, F. (1994). Autism: beyond "theory of mind." *Cognition* 50, 115–132. doi: 10.1016/0010-0277(94)90024-8

Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778. doi: 10.1038/nn1268

Hamilton, R. H., Shenton, J. T., and Coslett, H. B. (2006). An acquired deficit of audiovisual speech processing. *Brain Lang.* 98, 66–73. doi: 10.1016/j.bandl.2006.02.001

Happé, F. (1999). Autism: cognitive deficit or cognitive style? *Trends Cogn. Sci.* 3, 216–222. doi: 10.1016/S1364-6613(99)01318-2

Happé, F. (2005). "The weak central coherence account of autism," in *Handbook of Autism and Pervasive Developmental Disorders, 3rd Edn.*, Vol. 1, eds F. R. Volkmar, R. Paul, A. Klin, D. Cohen (Hoboken, NJ: John Wiley & Sons Inc.), 640–649.

Happé, F. G. (1994). Wechsler IQ profile and theory of mind in autism: a research note. *J. Child Psychol. Psychiatry* 35, 1461–1471. doi: 10.1111/j.1469-7610.1994.tb01287.x

Happé, F., and Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J. Autism Dev. Disord.* 36, 5–25. doi: 10.1007/s10803-005-0039-0

Hillock, A. R., Powers, A. R., and Wallace, M. T. (2011). Binding of sights and sounds: age-related changes in multisensory temporal processing. *Neuropsychologia* 49, 461–467. doi: 10.1016/j.neuropsychologia.2010.11.041

Hillock-Dunn, A., and Wallace, M. T. (2012). Developmental changes in the multisensory temporal binding window persist into

adolescence. *Dev. Sci.* 15, 688–696. doi: 10.1111/j.1467-7687.2012.01171.x

Iarocci, G., and McDonald, J. (2006). Sensory integration and the perceptual experience of persons with autism. *J. Autism Dev. Disord.* 36, 77–90. doi: 10.1007/s10803-005-0044-3

Irwin, J. R., Tornatore, L. A., Brancazio, L., and Whalen, D. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Dev.* 82, 1397–1403. doi: 10.1111/j.1467-8624.2011.01619.x

Jolliffe, T., and Baron-Cohen, S. (1999). A test of central coherence theory: linguistic processing in high-functioning adults with autism or Asperger syndrome: is local coherence impaired? *Cognition* 71, 149–185. doi: 10.1016/S0010-0277(99)00022-0

Jones, R., Quigney, C., and Huws, J. (2003). First-hand accounts of sensory perceptual experiences in autism: a qualitative analysis. *J. Intellect. Dev. Disabil.* 28, 112–121. doi: 10.1080/1366825031000147058

Just, M. A., Cherkassky, V. L., Keller, T. A., and Minshew, N. J. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain* 127, 1811–1821. doi: 10.1093/brain/awh199

Kanner, L. (1943). Autistic disturbances of affective contact. *Nerv. Child* 2, 217–250.

Kavanagh, J. F., and Mattingly, I. G. (1974). *Language by Ear and by Eye*. Boston, MA: MIT Press.

Kwakye, L. D., Foss-Feig, J. H., Cascio, C. J., Stone, W. L., and Wallace, M. T. (2011). Altered auditory and multisensory temporal processing in autism spectrum disorders. *Front. Integr. Neurosci.* 4:129. doi: 10.3389/fnint.2010.00129

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* 30, 205–223. doi: 10.1023/A:1005592401947

Markram, H., Rinaldi, T., and Markram, K. (2007). The intense world syndrome–an alternative hypothesis for autism. *Front. Neurosci.* 1:77. doi: 10.3389/neuro.01.1.1.006.2007

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Mongillo, E., Irwin, J., Whalen, D., Klaiman, C., Carter, A., and Schultz, R. (2008). Audiovisual processing in children with and without autism spectrum disorders. *J. Autism Dev. Disord.* 38, 1349–1358. doi: 10.1007/s10803-007-0521-y

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cogn. Psychol.* 9, 353–383. doi: 10.1016/0010-0285(77)90012-3

Noens, I. L., and Berckelaer-Onnes, I. A. V. (2005). Captured by details: sense-making, language and communication in autism. *J. Commun. Disord.* 38, 123–141. doi: 10.1016/j.jcomdis.2004.06.002

Pouget, P., Stepniewska, I., Crowder, E. A., Leslie, M. W., Emeric, E. E., Nelson, M. J., et al. (2009). Visual and motor connectivity and the distribution of calcium-binding proteins in macaque frontal eye field: implications for saccade target selection. *Front. Neuroanat.* 3:2. doi: 10.3389/neuro.05.002.2009

Powers, A. R. 3rd., Hillock, A. R., and Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *J. Neurosci.* 29, 12265–12274. doi: 10.1523/JNEUROSCI.3501-09.2009

Rippon, G., Brock, J., Brown, C., and Boucher, J. (2007). Disordered connectivity in the autistic brain: challenges for the "new psychophysiology." *Int. J. Psychophysiol.* 63, 164–172. doi: 10.1016/j.ijpsycho.2006.03.012

Rogers, S. J., and Vismara, L. A. (2008). Evidence-based comprehensive treatments for early autism. *J. Clin. Child Adolesc. Psychol.* 37, 8–38. doi: 10.1080/15374410701817808

Schlessinger, J. J., Stevenson, R. A., Shotwell, M. S., and Wallace, M. T. (in press). Improving pulse oximetry pitch perception with multisensory perceptual training. *Anesth. Anesthesiol.*

Smith, E. G., and Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *J. Child Psychol. Psychiatry* 48, 813–821. doi: 10.1111/j.1469-7610.2007.01766.x

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B., and James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *Neuroimage* 49, 3308–3318. doi: 10.1016/j.neuroimage.2009.12.001

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., et al. (2014). Multisensory temporal integration in autism spectrum disorders. *J. Neurosci.* 34, 691–697. doi: 10.1523/JNEUROSCI.3615-13.2014

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., et al. (in press). Brief report: arrested development of audiovisual speech perception in autism spectrum disorders. *J. Autism Dev. Disord.* 1–8. doi: 10.1007/s10803-013-1992-7

Stevenson, R. A., VanDerKlok, R. M., Pisoni, D. B., and James, T. W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *Neuroimage* 55, 1339–1345. doi: 10.1016/j.neuroimage.2010.12.063

Stevenson, R. A., Wilson, M. M., Powers, A. R., and Wallace, M. T. (2013). The effects of visual training on multisensory temporal processing. *Exp. Brain Res.* 225, 479–489. doi: 10.1007/s00221-012-3387-y

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Taylor, N., Isaac, C., and Milne, E. (2010). A comparison of the development of audiovisual integration in children with autism spectrum disorders and typically developing children. *J. Autism Dev. Disord.* 40, 1403–1411. doi: 10.1007/s10803-010-1000-4

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884. doi: 10.3758/APP.72.4.871

Williams, J., Massaro, D. W., Peel, N. J., Bosseler, A., and Suddendorf, T. (2004). Visual-auditory integration during speech imitation in autism. *Res. Dev. Disabil.* 25, 559–575. doi: 10.1016/j.ridd.2004.01.008

Wolery, M., and Garfinkle, A. N. (2002). Measures in intervention research with young children who have autism. *J. Autism Dev. Disord.* 32, 463–478. doi: 10.1023/A:1020598023809

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., and Wallace, M. T. (2013). Multisensory speech perception in children with autism spectrum disorders. *J. Autism Dev. Disord.* 43, 2891–2902. doi: 10.1007/s10803-013-1836-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 March 2014; accepted: 10 April 2014; published online: 21 May 2014.

Citation: Stevenson RA, Segers M, Ferber S, Barense MD and Wallace MT (2014) The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. Front. Psychol. ***5****:379. doi: 10.3389/fpsyg.2014.00379*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 10 July 2014
doi: 10.3389/fpsyg.2014.00725

What is the McGurk effect?

Kaisa Tiippana*

Division of Cognitive Psychology and Neuropsychology, Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland
**Correspondence: kaisa.tiippana@helsinki.fi*

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Lawrence Brancazio, Southern Connecticut State University, USA

Keywords: audiovisual, illusion, integration, McGurk effect, multisensory, perception, speech

McGurk and MacDonald (1976) reported a powerful multisensory illusion occurring with audiovisual speech. They recorded a voice articulating a consonant and dubbed it with a face articulating another consonant. Even though the acoustic speech signal was well recognized alone, it was heard as another consonant after dubbing with incongruent visual speech. The illusion has been termed the McGurk effect. It has been replicated many times, and it has sparked an abundance of research. The reason for the great impact is that this is a striking demonstration of multisensory integration. It shows that auditory and visual information is merged into a unified, integrated percept. It is a very useful research tool since the strength of the McGurk effect can be taken to reflect the strength of audiovisual integration.

Here I shall make two main claims regarding the definition and interpretation of the McGurk effect since they bear relevance to its use as a measure of multisensory integration. First, the McGurk effect should be defined as a categorical change in auditory perception induced by incongruent visual speech, resulting in a single percept of hearing something other than what the voice is saying. Second, when interpreting the McGurk effect, it is crucial to take into account the perception of the unisensory acoustic and visual stimulus components.

There are many variants of the McGurk effect (McGurk and MacDonald, 1976; MacDonald and McGurk, 1978)[1]. The best-known case is when dubbing a voice saying [b] onto a face articulating [g] results in hearing [d]. This is called the fusion effect since the percept differs from the acoustic and visual components. Many researchers have defined the McGurk effect exclusively as the fusion effect because here integration results in the perception of a third consonant, obviously merging information from audition and vision (van Wassenhove et al., 2007; Keil et al., 2012; Setti et al., 2013). This definition ignores the fact that other incongruent audiovisual stimuli produce different types of percepts. For example, a reverse combination of these consonants, A[g]V[b], is heard as [bg], i.e., the visual and auditory components one after the other. There are other pairings, which result in hearing according to the visual component, e.g., acoustic [b] presented with visual [d] is heard as [d]. Here my first claim is that the definition of the McGurk effect should be that an acoustic utterance is heard as another utterance when presented with discrepant visual articulation. This definition includes all variants of the illusion, and it has been used by MacDonald and McGurk (1978) themselves, as well as by several others (e.g., Rosenblum and Saldaña, 1996; Brancazio et al., 2003). The different variants of the McGurk effect represent the outcome of audiovisual integration. When integration takes place, it results in a unified percept, without access to the individual components that contributed to the percept. Thus, when the McGurk effect occurs, the observer has the subjective experience of hearing a certain utterance, even though another utterance is presented acoustically.

One challenge with this interpretation of the McGurk effect is that it is impossible to be certain that the responses the observer gives correspond to the actual percepts. The real McGurk effect arises due to multisensory integration, resulting in an altered auditory percept. However, if integration does not occur, the observer can perceive the components separately and may choose to respond either according to what he heard or according to what he saw. This is one reason why the fusion effect is so attractive: If the observer reports a percept that differs from both stimulus components, he does not seem to rely on either modality alone, but instead really fuse the information from both. However, this approach does not guarantee a straightforward measure of integration any more than the other variants of the illusion, as is argued below.

The second main claim here is that the perception of the acoustic and visual stimulus components has to be taken into account when interpreting the McGurk effect. This issue has been elaborated previously in the extensive work by Massaro and colleagues (Massaro, 1998) and others (Sekiyama and Tohkura, 1991; Green and Norrix, 1997; Jiang and Bernstein, 2011). It is important because the identification accuracy of unisensory components is reflected into audiovisual speech perception.

In general, the strength of the McGurk effect is taken to increase when the proportion of responses according to the acoustic component decreases and/or when the proportion of fusion responses increases. That is, the McGurk effect for stimulus A[b]V[g] is considered stronger when fewer B responses and/or more D responses are given. This is often an adequate way to measure the strength of the McGurk effect—if one keeps in mind that

[1]Throughout this paper only some representative references are mentioned as examples of the extensive literature on each topic.

it implicitly assumes that perception of the acoustic and visual components is accurate (or at least constant across conditions that are compared). However, it can lead to erroneous conclusions if this assumption does not hold.

The fusion effect provides a prime example of this caveat. It has been interpreted to mean that acoustic and visual information is integrated to produce a novel, intermediate percept. For example, when A[b]V[g] is heard as [d], the percept is thought to emerge due to fusion of the features (for the place of articulation) provided via audition (bilabial) and vision (velar), so that a different, intermediate consonant (alveolar) is perceived (van Wassenhove, 2013). However, already McGurk and MacDonald (1976) themselves wrote that "lip movements for [ga] are frequently misread as [da]," even though they did not measure speechreading performance, unfortunately. The omission of the unisensory visual condition in the original study is one factor that has contributed to the strong status of the fusion effect as the only real McGurk effect, reflecting true integration. Still, if visual [g] is confused with [d], it is not at all surprising or special if A[b]V[g] is perceived as [d].

To demonstrate the contribution of the unisensory components more explicitly, I'll take two examples of my research, in which fusion-type stimuli produced different percepts depending on the clarity of the visual component. In one study, a McGurk stimulus A[epe]V[eke] was mainly heard as a fusion [ete] (Tiippana et al., 2004). This reflected the fact that in a visual-only identification task, the visual [eke] was confused with [ete] (42% K responses and 45% T responses to visual [eke]). In another study, a McGurk stimulus A[apa]V[aka] was mainly heard as [aka], and this could be traced back to the fact that in a visual-only identification task, the visual [aka] was clearly distinguishable from [ata], and thus recognized very accurately (100% correct in typical adults; Saalasti et al., 2012; but note the deviant behavior of individuals with Asperger syndrome). Thus, even though the McGurk stimuli were of a fusion type in both studies, their perception differed depending largely on the clarity of the visual components. These findings underscore the importance of knowing the perceptual qualities of the unisensory stimuli before making conclusions about multisensory integration.

Exactly how to take the properties of the unisensory components into account in multisensory perception of speech is beyond this paper. Addressing this issue in detail requires carefully designed experimental studies (Bertelson et al., 2003; Alsius et al., 2005), computational modeling (Massaro, 1998; Schwartz, 2010), and investigation of the underlying brain mechanisms (Sams et al., 1991; Skipper et al., 2007). However, the main guideline is that unisensory perception of stimulus components is reflected into multisensory perception of the whole (Ernst and Bülthoff, 2004).

During experiments, when the task is to report what was heard, the observer reports the conscious auditory percept evoked by the audiovisual stimulus. If there is no multisensory integration or interaction, the percept is identical for the audiovisual stimulus and the auditory component presented alone. If there is audiovisual integration, the conscious auditory percept changes. To which extent visual input influences the percept depends on how coherent and reliable information each modality provides. Coherent information is integrated and weighted e.g., according to the reliability of each modality, which is reflected in unisensory discriminability.

This perceptual process is the same for audiovisual speech—be it natural, congruent audiovisual speech or artificial, incongruent McGurk speech stimuli. The outcome is the conscious auditory percept. Depending on the relative weighting of audition and vision, the outcome for McGurk stimuli can range from hearing according to the acoustic component (when audition is more reliable than vision) to fusion and combination percepts (when both modalities are informative to some extent) to hearing according to the visual component (when vision is more reliable than audition). Congruent audiovisual speech is treated no differently, showing visual influence when the auditory reliability decreases. The different variants of the McGurk effect are all results of this same perceptual process and reflect audiovisual integration.

The McGurk effect is an excellent tool to investigate multisensory integration in speech perception. The main messages of this opinion paper are, first, that the McGurk effect should be defined as a change in auditory perception due to incongruent visual speech, so that observers hear another speech sound than what the voice uttered, and second, that the perceptual properties of the acoustic and visual stimulus components should be taken into account when interpreting the McGurk effect as reflecting integration.

ACKNOWLEDGMENT

This research was funded by a grant from the University of Helsinki.

REFERENCES

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Bertelson, P., Vroomen, J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* 14, 592–597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Brancazio, L., Miller, J. L., and Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Percept. Psychophys.* 65, 591–601. doi: 10.3758/BF03194585

Ernst, M. O., and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169. doi: 10.1016/j.tics.2004.02.002

Green, K. P., and Norrix, L. W. (1997). Acoustic cues to place of articulation and the McGurk effect: the role of release bursts, aspiration, and formant transitions. *J. Speech Lang. Hear. Res.* 40, 646–665.

Jiang, J., and Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1193–1209. doi: 10.1037/a0023100

Keil, J., Muller, N., Ihssen, N., and Weisz, N. (2012). On the variability of the McGurk effect: audiovisual integration depends on prestimulus brain states. *Cereb. Cortex* 22, 221–231. doi: 10.1093/cercor/bhr125

MacDonald, J., and McGurk, H. (1978). Visual influences on speech perception processes. *Percept. Psychophys.* 24, 253–257. doi: 10.3758/BF03206096

Massaro, D. W. (1998). *Perceiving Talking Faces*. Cambridge, MA: MIT Press.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Rosenblum, L. D., and Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 22, 318–331. doi: 10.1037/0096-1523.22.2.318

Saalasti, S., Kätsyri, J., Tiippana, K., Laine-Hernandez, M., von Wendt, L., and Sams, M. (2012).

Audiovisual speech perception and eye gaze behavior of adults with Asperger Syndrome. *J. Autism Dev. Disord.* 42, 1606–1615. doi: 10.1007/s10803-011-1400-0

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-F

Schwartz, J. L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *J. Acoust. Soc. Am.* 127, 1584–1594. doi: 10.1121/1.3293001

Sekiyama, K., and Tohkura, Y. (1991). McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc. Am.* 90, 1797–1805.

Setti, A., Burke, K. E., Kenny, R., and Newell, F. N. (2013). Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes. *Front. Psychol.* 4:575. doi: 10.3389/fpsyg.2013.00575

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cereb. Cortex* 17, 2387–2399. doi: 10.1093/cercor/bhl147

Tiippana, K., Andersen, T. S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Front. Psychol.* 4:388. doi: 10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 March 2014; accepted: 23 June 2014; published online: 10 July 2014.
*Citation: Tiippana K (2014) What is the McGurk effect? Front. Psychol. **5**:725. doi: 10.3389/fpsyg.2014.00725*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 13 May 2014
doi: 10.3389/fpsyg.2014.00420

The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception

Avril Treille*, Coriandre Vilain and Marc Sato

CNRS, Département Parole and Cognition, Gipsa-Lab, UMR 5216, Grenoble Université, Grenoble, France

Edited by:
Riikka Mottonen, University of Oxford, UK

Reviewed by:
Joana Acha, Basque Centre on Cognition, Brain and Language, Spain
Takayuki Ito, Haskins Laboratories, USA

****Correspondence:***
Avril Treille, CNRS, Département Parole and Cognition, Gipsa-Lab, UMR 5216, Grenoble Université, 1180 Avenue Centrale, BP 25, 38040 Grenoble Cedex 9, France
e-mail: avril.treille@gipsa-lab.inpg.fr

Recent magneto-encephalographic and electro-encephalographic studies provide evidence for cross-modal integration during audio-visual and audio-haptic speech perception, with speech gestures viewed or felt from manual tactile contact with the speaker's face. Given the temporal precedence of the haptic and visual signals on the acoustic signal in these studies, the observed modulation of N1/P2 auditory evoked responses during bimodal compared to unimodal speech perception suggest that relevant and predictive visual and haptic cues may facilitate auditory speech processing. To further investigate this hypothesis, auditory evoked potentials were here compared during auditory-only, audio-visual and audio-haptic speech perception in live dyadic interactions between a listener and a speaker. In line with previous studies, auditory evoked potentials were attenuated and speeded up during both audio-haptic and audio-visual compared to auditory speech perception. Importantly, the observed latency and amplitude reduction did not significantly depend on the degree of visual and haptic recognition of the speech targets. Altogether, these results further demonstrate cross-modal interactions between the auditory, visual and haptic speech signals. Although they do not contradict the hypothesis that visual and haptic sensory inputs convey predictive information with respect to the incoming auditory speech input, these results suggest that, at least in live conversational interactions, systematic conclusions on sensory predictability in bimodal speech integration have to be taken with caution, with the extraction of predictive cues likely depending on the variability of the speech stimuli.

Keywords: audio-visual speech perception, audio-haptic speech perception, multisensory interactions, EEG, auditory evoked potentials

INTRODUCTION

How information from different sensory modalities, such as sight, sound and touch, is combined to form a single coherent percept? As central to adaptive behavior, multisensory integration occurs in everyday life when natural events in the physical world have to be integrated from different sensory sources. It is an highly complex process known to depend on the temporal, spatial and causal relationships between the sensory signals, to take place at different timescales in several subcortical and cortical structures and to be mediated by both feedforward and backward neural projections. In addition to their coherence, the perceptual saliency and relevance of each sensory signal from the external environment, as well as their predictability and joint probability to occur, also act on the integration process and on the representational format at which the sensory modalities interface (for reviews, see Stein and Meredith, 1993; Stein, 2012).

Audio-visual speech perception is a special case of multisensory processing that interfaces with the linguistic system. Although one can extract phonetic features from the acoustic signal alone, adding visual speech information from the speaker's face is known to improve speech intelligibility in case of a degraded acoustic signal (Sumby and Pollack, 1954; Benoît et al., 1994; Schwartz et al., 2004), to facilitate the understanding of a semantically complex statement (Reisberg et al., 1987) or a foreign language (Navarra and Soto-Faraco, 2005), and to benefit hearing-impaired listeners (Grant et al., 1998). Conversely, in laboratory settings, adding incongruent visual speech information may interfere with auditory speech perception and even create an illusory percept (McGurk and MacDonald, 1976). Finally, as in other cases of bimodal integration, audio-visual speech integration depends on the perceptual saliency of both the auditory (Green, 1998) and visual (Campbell and Massaro, 1997) speech signals, as well as their spatial (Jones and Munhall, 1997) and temporal (van Wassenhove et al., 2003) relationships.

At the brain level, several magneto-encephalographic (MEG) and electro-encephalographic (EEG) studies demonstrate that visual speech input modulates auditory activity as early as 50–100 ms in the primary and secondary auditory cortices (Sams et al., 1991; Klucharev et al., 2003; Lebib et al., 2003; Besle et al., 2004; Hertrich et al., 2007; Winneke and Phillips, 2011). Importantly, it has been shown that both the latency and amplitude of auditory evoked responses (N1/P2, M100) are attenuated and speeded up during audio-visual compared to auditory-only speech perception (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al.,

2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Pilling, 2010; Vroomen and Stekelenburg, 2010; Baart et al., 2014; Treille et al., 2014). Moreover, N1/P2 latency facilitation also appears to be directly function of the visemic information, with the higher visual recognition of the syllable, the longer latency facilitation (van Wassenhove et al., 2005; Arnal et al., 2009). Since the visual speech signal preceded the acoustic speech signal by 10s or 100s of milliseconds in these studies, the observed speeding-up and amplitude suppression of auditory evoked potentials might both reflect non-speech specific temporal (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010) and phonetic (van Wassenhove et al., 2005; Arnal et al., 2009) visual predictions of the incoming auditory syllable (for recent discussions, see Arnal and Giraud, 2012; van Wassenhove, 2013; Baart et al., 2014).

Interestingly, speech can be perceived not only by the ear and by the eye but also by the hand, with orofacial speech gestures felt and monitored from manual tactile contact with the speaker's face. Past studies on the Tadoma method provide evidence for successful communication abilities in trained deaf-blind individuals through the haptic modality (Alcorn, 1932; Norton et al., 1977). A few behavioral studies also demonstrate the influence of tactile information on auditory speech perception in untrained individuals without sensory impairment, especially in case of noisy or ambiguous acoustic signals (Fowler and Dekle, 1991; Gick et al., 2008; Sato et al., 2010). In a recent EEG study (Treille et al., 2014), electrophysiological evidence of cross-modal interactions was found during both audio-visual and audio-haptic speech perception, through the course of live dyadic interactions between a listener and a speaker. In this study, participants were seated at arm's length from an experimenter and they were instructed to manually categorize /pa/ or /ta/ syllables presented auditorily, visually and/or haptically. In line with the above-mentioned EEG/MEG studies, N1 auditory evoked responses were attenuated and speeded up during live audio-visual speech perception. Crucially, haptic information was also found to speed up auditory speech processing as early as 100 ms. Given the temporal precedence of the dynamic configurations of the articulators on the auditory signal, as attested in a behavioral control experiment, the observed audio-haptic interactions in the listener's brain raise the possibility that the brain use predictive temporal and/or phonetic relevant tactile information for auditory processing, despite less natural processing to extract relevant speech information from the haptic modality. From this possibility, however, a clear limit of this study comes from the use of a simple two-alternative forced-choice identification task between /pa/ and /ta/ syllables and an insufficient number of trials for reliable EEG analyses per syllable.

To further explore whether perceivers might integrate tactile information in auditory speech perception as they do with visual information, the present study aimed at replicating the observed bimodal interactions during live face-to-face and hand-to-face speech perception (Treille et al., 2014). As observed in previous studies on audio-visual speech perception (van Wassenhove et al., 2005; Arnal et al., 2009), we also specifically tested whether modulation of N1/P2 auditory evoked potentials during both audio-visual and audio-haptic speech perception might depend on the degree to which the haptic and visual signals predict the incoming auditory speech target. To this aim, the experimental procedure was adapted from the Tadoma method and similar to that previously used by Treille et al. (2014), except the use of a three-alternative forced-choice identification task between /pa/, /ta/, and /ka/ syllables and a sufficient number of trials for reliable EEG analyses per syllable. A gradient of visual and haptic recognition between the three syllables was first attested in a behavioral experiment, which was a requirement to assess visual and haptic predictability on the incoming auditory signal in a subsequent EEG experiment. In line with previous EEG studies on audio-visual speech integration (van Wassenhove et al., 2005; Arnal et al., 2009), we hypothesized that the higher visual and haptic recognition of the syllable, the stronger latency facilitation in the audio-visual and audio-haptic modalities.

MATERIALS AND METHODS

PARTICIPANTS

Sixteen healthy adults, native French speakers, participated in the study (eight females; mean age ± SD, 29 ± 8 years). All participants were right-handed, had normal or corrected-to-normal vision and reported no history of speaking, hearing or motor disorders. Written informed consent was obtained for all participants and they were compensated for the time spent in the study. The study was approved by the Grenoble University Ethical Committee.

STIMULI

Based on a previous EEG study (van Wassenhove et al., 2005), /pa/, /ta/, and /ka/ syllables were selected in order to ensure precise acoustic onsets (thanks to the unvoiced stop bilabial /p/, alveolar /t/, and velar /k/ stop consonants) crucial for EEG analyses and, importantly, to ensure a gradient of visual and haptic recognition between these syllables (with notably the bilabial /p/ consonant known to be more visually salient than alveolar /t/ and velar /k/ consonants).

EXPERIMENTAL PROCEDURE

The study consisted on one behavioral experiment immediately followed by one EEG experiment. The behavioral experiment was performed in order to ensure a gradient of visual and haptic recognition of /pa/, /ta/, and /ka/ syllables. Importantly, since individual syllable onsets of the experimenter's productions were used as acoustical triggers for EEG analyses, the visual and haptic modalities of presentation were not included in the EEG experiment. In both experiments, Presentation software (Neurobehavioral Systems, Albany, CA, USA) was used to control the visual stimuli for the experimenter, the audio stimuli (beep) for the participant and to record key responses. In addition, all experimenter productions were recorded for off-line analyses in the EEG experiment.

Behavioral experiment

In a first behavioral experiment, participants were individually tested in a sound-proof room and were seated at arm's length from a female experimenter (see **Figure 1A**).

They were told that they would be presented with /pa/, /ta/, or /ka/ syllables either auditorily, visually, audio-visually, haptically, or audio-haptically over the hand-face contact. In the auditory modality (A), participants were instructed to keep their eyes closed and to listen to each syllable overtly produced by the

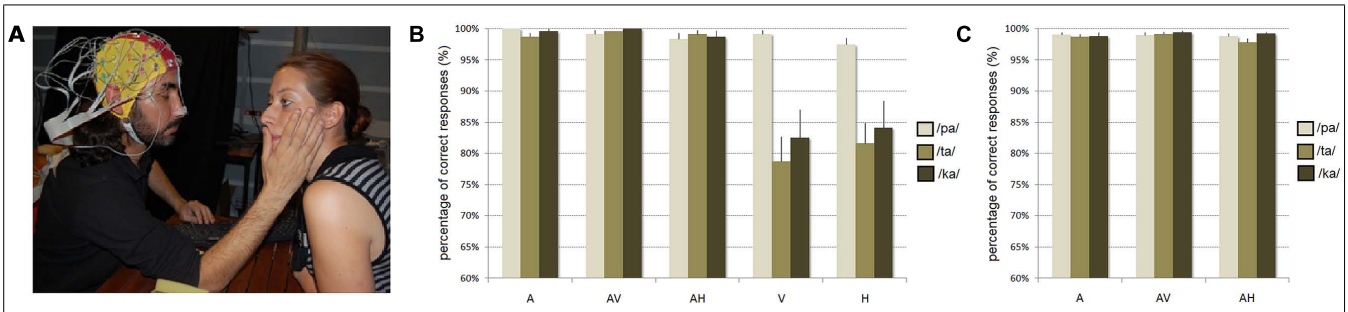


FIGURE 1 | (A) Experimental design used in the audio-haptic (AH) modality. In the haptic (H) and AH modalities, participants were asked to keep their eyes closed with their right hand placed on the experimenter's face and to categorize with their left hand each perceived syllable. In the auditory modality (A), participants were instructed to keep their eyes closed while, in the visual (V) and audio-visual modality (AV), they were asked to also look at the experimenter's face. The behavioral experiment included A, V, H, AV, AH modalities while the EEG experiment only included A, AV, and AH modalities. **(B,C)** Mean percentage of correct identification for /pa/, /ta/, and /ka/ syllables in each modality of presentation in the **(B)** behavioral and **(C)** EEG experiments. Error bars represent standard errors of the mean.

experimenter. In the audio-visual modality (AV), they were asked to also look at the experimenter's face. In the audio-haptic modality (AH), they were asked to keep their eyes closed with their right hand placed on the experimenter's face (the thumb placed lightly and vertically against the experimenter's lips and the other fingers placed horizontally along the jaw line in order to help distinguishing both lip and jaw movements). This experimental procedure was adapted from the Tadoma method and similar to that previously used by Treille et al. (2014). Finally, the visual-only (V) and haptic-only (H) modalities were similar to the AV and AH modalities except that the experimenter silently produced each syllable.

The experimenter faced the participant and a computer screen placed behind the participant. On each trial, the computer screen specified the syllable to be produced. To this aim, the syllable was printed three times on the computer screen at 1 Hz, with the last display serving as the visual go-signal to produce the syllable. The inter-trial interval was 3 s. The experimenter previously practiced and learned to articulate each syllable in synchrony with the visual go-signal, with an initial neutral closed-mouth position and maintaining an even intonation, tempo and vocal intensity.

A three-alternative forced-choice identification task was used, with participants instructed to categorize each perceived syllable by pressing on one of three keys corresponding to /pa/, /ta/, or /ka/ on a computer keyboard with their left hand. A brief single audio beep was delivered 600 ms after the visual go-signal (expecting to occur in synchrony with the experimenter production) with the participants told to produce their responses only after this audio go-signal. This procedure was done in order to dissociate sensory/perceptual responses from motor responses on EEG data in the next experiment. As a consequence, no reaction-times were acquired and only response rate were considered in further analyses.

Every syllable (/pa/, /ta/, or /ka/) was presented 15 times in each modality (A, V, H, AV, AH) in a single randomized sequence for a total of 225 trials. The response key designation were counterbalanced across participants. Before the experiment, participants performed few practice trials in all modalities. They received no instructions concerning how to interpret visual and haptic information but they were asked to pay attention to both modalities during bimodal presentation.

EEG experiment

Because of no possible reliable acoustical triggers in the visual-only and haptic-only modalities, the EEG experiment only included three individual experimental sessions related to A, AV, and AH modalities of presentation. Except this difference and the number of trials, the experimental procedure was identical to that used in the behavioral experiment. In each session, every syllable (/pa/, /ta/, or /ka/) was presented 80 times in a randomized sequence for a total of 240 trials. The order of the modality of presentation and the response key designation were fully counterbalanced across participants. Because the experimental procedure was quite taxing, each experimental session was split into two blocks of around 6 min each, allowing short breaks for both the experimenter and the participants.

EEG ACQUISITION

In the EEG experiment, EEG data were continuously recorded from 64 scalp electrodes (Electro-Cap International, INC., according to the international 10–20 system) using the Biosemi ActiveTwo AD-box EEG system operating at a sampling rate of 256 Hz. Two additional electrodes served as reference (common mode sense [CMS] active electrode) and ground (driven right leg [DRL] passive electrode). One other external reference electrode was at the top of the nose. The electro-oculogram measuring horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes at the outer canthus of each eye as well as above and below the right eye. Before the experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

DATA ANALYSES

Behavioral analyses

In both the behavioral and EEG experiments, the proportion of correct responses was individually determined for each participant, each syllable and each modality. Two-way repeated-measure ANOVAs were performed on these data with the modality

(A, V, H, AV, AH in the behavioral experiment; A, AV, AH in the EEG experiment) and the syllable (/pa, /ta/, /ka/) as within-subjects variables.

Acoustical analyses

In the EEG experiment, acoustical analyses were performed on the experimenter's recorded syllables in order to determine the individual syllable onsets serving as acoustical triggers for the EEG analyses. All acoustical analyses were performed using Praat software (Boersma and Weenink, 2013). First, an automatic procedure based on an intensity and duration algorithm detection roughly identified each syllable's onset in the A, AV, and AH modalities (11520 utterances). For all syllables, these onsets were further manually and precisely determined, based on waveform and spectrogram information related to the acoustic characteristics of voiced stop consonants. Omissions and wrong productions were identified and removed from the analyses (less than 1%).

EEG analyses

EEG data were processed using the EEGLAB toolbox (Delorme and Makeig, 2004) running on Matlab (Mathworks, Natick, MA, USA). Since N1/P2 auditory evoked potentials have maximal response over central sites on the scalp (Scherg and Von Cramon, 1986; Näätänen and Picton, 1987), EEG data preprocessing and analyses were conducted on three central electrodes (C3, Cz, C4). These electrodes, covering left, middle, and right central sites, were also selected based on previous EEG studies on audio-visual speech perception (e.g., Klucharev et al., 2003; Besle et al., 2004; Pilling, 2010; Treille et al., 2014). EEG data were first re-referenced offline to the nose recording and band-pass filtered using a two-way least-squares FIR filtering (1–20 Hz). Data were then segmented into epochs of 1000 ms (from −500 ms to +500 ms to the acoustic syllable onset, individually determined from the acoustical analyses), with the prestimulus baseline defined from −500 ms to −400 ms. Epochs with an amplitude change exceeding ±60 μV at any channel (including HEOG and VEOG channels) were rejected (on average, less than 10%).

For each participant and each modality, the peak latency of auditory N1 and P2 evoked responses were first determined on the EEG waveform averaged over all electrodes and syllables. For each syllable, two temporal windows were then defined on these peaks ±30 ms in order to individually calculate N1 and P2 amplitude and latency on the related average waveform of C3, Cz, C4 electrodes. Two-way repeated-measure ANOVAs were then performed on N1 and P2 amplitude and latency with the modality (A, AV, AH) and the syllable (/pa/, /ka/, /ta/) as within-subjects variables.

In order to confirm previous EEG/MEG studies demonstrating that P2 and M100 latency reduction in the audio-visual modality vary as a function of the visual recognition of the presented syllable (van Wassenhove et al., 2005; Arnal et al., 2009), additional Pearson's correlation analyses were carried out. These correlation analyses were performed between the individual visual and haptic recognition scores of the three syllables in the behavioral experiment and the related latency facilitation and reduction amplitude observed in the AV and AH modalities in the EEG experiment (leading to 3 × 16 correlation points per measure and per modality). In addition to raw data, these analyses were also performed on individual Z-score normalized data, in order to take account of individual differences.

RESULTS

For all the following analyses, the significance level was set at $p = 0.05$ and Greenhouse–Geisser corrected (for violation of the sphericity assumption) when appropriate. When required, *post hoc* analyses were conducted with Newman–Keuls tests.

BEHAVIORAL ANALYSES

Behavioral experiment (see Figure 1B)

Overall, the mean proportion of correct responses was of 94%. The main effect of modality of presentation was significant [$F(4,60) = 33.67$, $p < 0.001$], with more correct responses in A, AV, and AH modalities than in V and H modalities (as shown by *post hoc* analyses, all p's < 0.001). Significant differences were also observed between syllables [$F(2,30) = 15.59$, $p < 0.001$], with more correct responses for /pa/ than for /ta/ and /ka/ syllables (as shown by *post hoc* analyses, all p's < 0.001). Finally, the interaction between the modality and the syllable was also reliable [$F(8,120) = 7.39$, $p < 0.001$]. While no significant differences were observed between syllables in A, AV, and AH modalities (with almost perfect identification for all syllables), more correct responses were observed for /pa/ than for /ta/ and /ka/ syllables in both V and H modalities (as shown by *post hoc* analyses, all p's < 0.001). Altogether, these results thus demonstrate a near perfect identification of /pa/ in all modalities, but a lower accuracy for /ta/ and /ka/ syllables in V and H modalities.

EEG experiment (see Figure 1C)

In the EEG experiment, the mean proportion of correct responses was of 99%. No significant effect of the modality [$F(2,30) = 1.72$], syllable [$F(2,30) = 1.34$] or interaction [$F(4,60) = 0.90$] was observed, with a near perfect identification of all syllables in A, AV, and AH modalities.

EEG ANALYSES

N1 amplitude (see Figures 2 and 3A-left)

The main effect of modality was significant [$F(2,30) = 9.19$, $p < 0.001$], with a reduced negative N1 amplitude observed in the AV and AH modalities as compared to the A modality (as shown by *post hoc* analyses, $p < 0.001$ and $p < 0.02$, respectively; on average, A: −5.3 μV, AV: −3.1 μV, AH: −4.1 μV). The interaction between the modality and the syllable was also found to be significant [$F(4,60) = 7.23$, $p < 0.001$]. While for /pa/ a significant amplitude reduction was observed in both AV and AH modalities as compared to the A modality, an amplitude reduction was only observed in the AV modality for /ta/ and /ka/ syllables (as shown by *post hoc* analyses, all p's < 0.001, see **Figure 3A**-left). In sum, these results demonstrate a visually induced amplitude suppression for all syllables and, importantly, an haptically induced amplitude suppression but only for /pa/ syllable.

P2 amplitude (see Figures 2 and 3B-left)

No significant effect of the modality [$F(2,30) = 1.91$], the syllable [$F(2,30) = 1.09$] and their interaction [$F(4,60) = 1.58$] was observed.

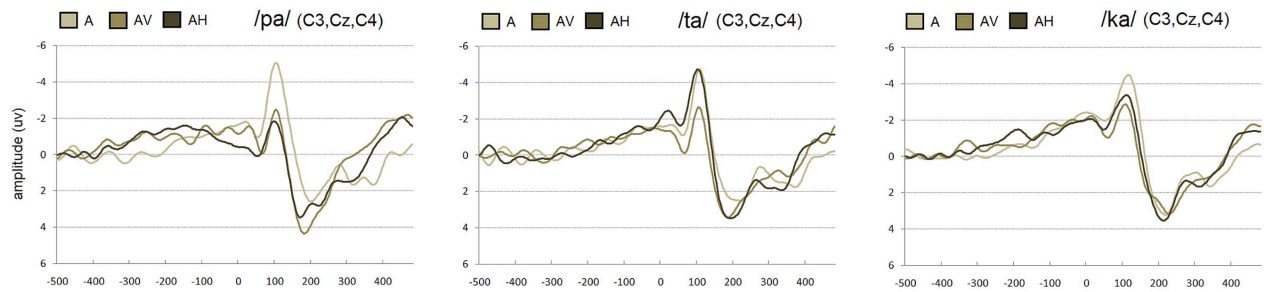


FIGURE 2 | Grand-average of auditory evoked potentials for /pa/, /ta/, and /ka/ syllables averaged over the left (C3), middle (Cz), and right (C4) central electrodes in the auditory, audio-visual, and audio-haptic modalities.

N1 latency (see Figures 2 and 3C-left)

No significant effect of the modality [$F(2,30) = 0.36$], the syllable [$F(2,30) = 3.13$] and their interaction [$F(4,60) = 1.78$] was observed.

P2 latency (see Figures 2 and 3D-left)

The main effect of syllable [$F(2,30) = 4.54$, $p < 0.02$] was reliable, with shorter P2 latencies observed for /pa/ and /ta/ syllables as compared to /ka/ (as shown by *post hoc* analyses, all p's < 0.03; on average, /pa/: 210 ms, /ta/: 211 ms, /ka/: 217 ms). Crucially, the main effect of modality was significant [$F(2,30) = 4.05$, $p < 0.03$], with shorter latencies in AV and AH as compared to the A modality (as shown by *post hoc* analyses, all p's < 0.05; on average, A: 223 ms, AV: 208 ms, AH: 207 ms). In sum, these results thus indicate faster processing of the P2 auditory evoked potential for /pa/ and /ka/ syllables. In addition, a latency facilitation was observed in both AV and AH modalities, irrespective of the presented syllables.

Correlation between perceptual recognition scores (see Figure 3-right)

For raw data, whatever the modality, no significant correlation was however observed for both N1 amplitude (AV: $r = 0.09$, $p = 0.54$; AH: $r = 0.06$, $p = 0.70$), P2 amplitude (AV: $r = 0.25$, $p = 0.09$; AH: $r = -0.09$, $p = 0.53$), N1 latency (AV: $r = -0.06$, $p = 0.71$; AH: $r = 0.11$, $p = 0.45$), and P2 latency (AV: $r = 0.07$, $p = 0.66$; AH: $r = -0.01$, $p = 0.92$). Results on additional correlation analyses on normalized data also failed to demonstrate any significant correlation for both N1 and P2 amplitude (N1-AV: $r = 0.01$, $p = 0.98$; N1-AH: $r = 0.18$, $p = 0.87$; P2-AV: $r = 0.21$, $p = 0.15$; P2-AH: $r = 0.02$, $p = 0.91$) and latency (N1-AV: $r = 0.01$, $p = 0.92$; N1-AH: $r = 0.12$, $p = 0.65$; P2-AV: $r = 0.06$, $p = 0.68$; P2-AH: $r = -0.02$, $p = 0.87$).

DISCUSSION

Two main results emerge from the present study. First, in line with our previous results (Treille et al., 2014), a modulation of N1/P2 auditory evoked potentials was observed during live audio-visual and audio-haptic speech perception compared to auditory speech perception. However, contrary to two previous studies of audio-visual speech perception (van Wassenhove et al., 2005; Arnal et al., 2009), no significant correlation was observed between the latency facilitation observed in the bimodal conditions and the degree of visual and haptic recognition of the presented syllables.

Before we discuss these results, it is first important to consider one potential limitation of the present study. Classically, testing cross-modal interactions requires to determine that the observed response in the bimodal condition differ to the sum of those observed in the unimodal conditions (e.g., AV $\neq$ A + V). However, visual-only and haptic-only modalities were not here tested, due to the technical difficulty to get temporal accurate and reliable triggers for EEG analyses. Notably, because of their temporal limitation and variability, visual and/or surface electromyographic recordings of the experimenter's lip, jaw or tongue movements would not allowed to determine reliable triggers (especially in the case of lip stretching for /ta/ and /ka/ syllables). From the possibility that the observed bimodal neural responses simply come from a superposition of the unimodal signals, it should however be noted that auditory evoked potentials are rarely observed in the visual-only modality in central electrodes (Besle et al., 2004; van Wassenhove et al., 2005; Pilling, 2010). Furthermore, in our previous study and using the same experimental design, we obtained behavioral evidence for a strong temporal precedence of the haptic and visual signals on the acoustic signal (Treille et al., 2014). In our view, it is therefore unlikely that visual and haptic event-related potentials might arise at the same time-latency and at the same central electrodes that N1 and P2 auditory evoked potentials. For these reasons, we here compared neural responses in each bimodal condition to the related unimodal condition (i.e., AV $\neq$ A and AH $\neq$ H), a testing procedure that has previously demonstrated latency facilitation and amplitude reduction of auditory evoked potentials in audio-visual compared to auditory-only speech perception (van Wassenhove et al., 2005; Pilling, 2010).

In spite of this limitation, the observed modulation of N1/P2 auditory evoked potentials in the audio-visual condition strongly suggests cross-modal speech interactions. It is first worthwhile noting that, for each participant, the three syllables were randomly presented in each session in order to minimize repetition effects, and the order of the modality of presentation was fully counterbalanced across participants so that possible overlapping modality effects are unlikely. In addition, auditory-evoked responses were compared between modalities, with the same number of trials and therefore similar possible habituation effects. Although our results

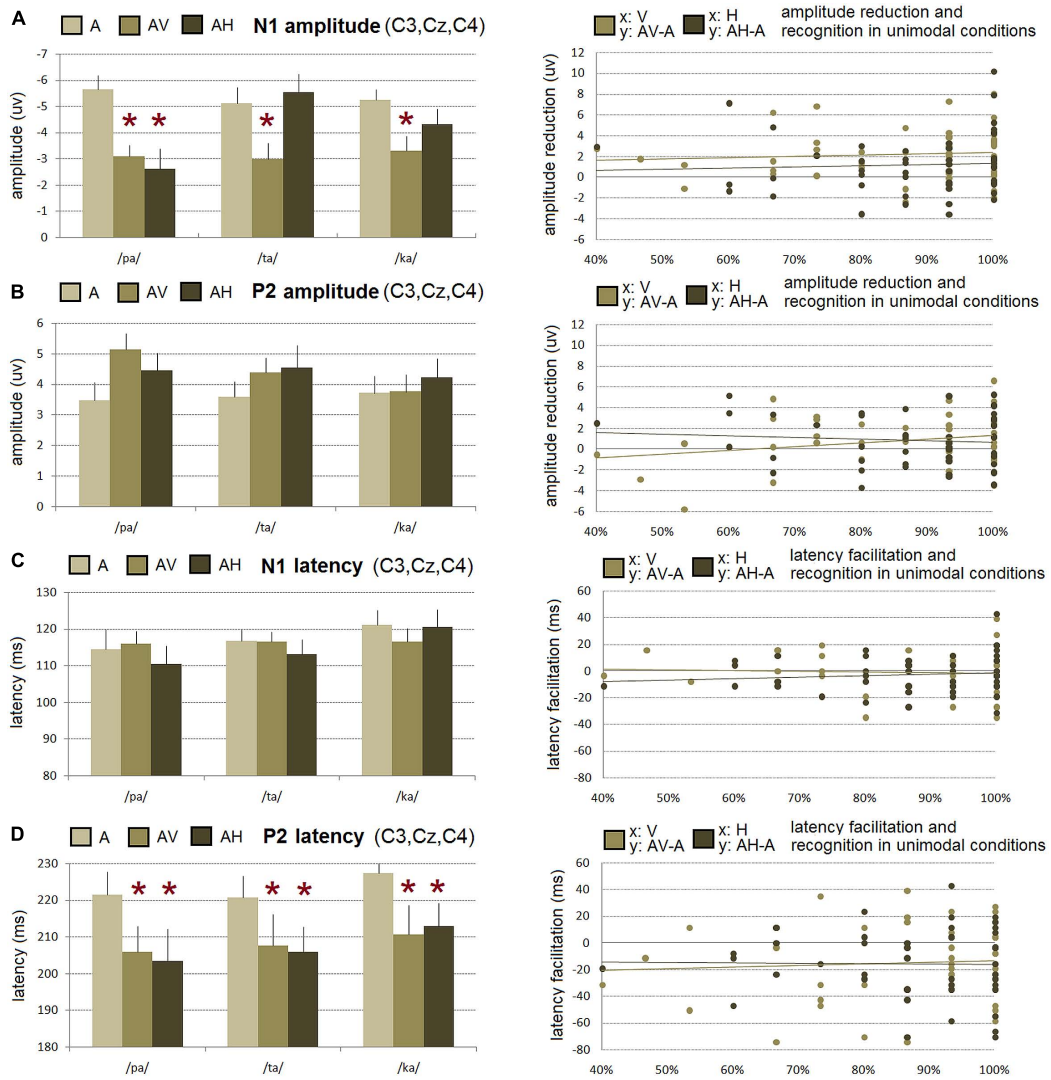


FIGURE 3 | Left. Mean N1 **(A)** and P2 **(B)** amplitude and mean N1 **(C)** and P2 **(D)** latency for /pa/, /ta/, and /ka/ syllables averaged over left (C3), middle (Cz), and right (C4) central electrodes in the auditory (A), audio-visual (AV), and audio-haptic (AH) modalities. Error bars represent standard errors of the mean. * indicates a significant effect.

Right. Correlation on raw data between the recognition scores observed in the visual-only and haptic-only modalities in the behavioral experiment (*x*-axis) and the reduction amplitude and latency facilitation observed in the audio-visual and audio-haptic modalities in the EEG experiment (*y*-axis). No correlation was significant.

appear globally consistent with previous EEG studies, some differences have however to be mentioned. First, while the observed amplitude reduction was here confined to the N1 auditory evoked potential, as in our previous study (Treille et al., 2014; see also Besle et al., 2004), such a visually induced suppression has been previously observed for both N1 and P2 auditory components (Klucharev et al., 2003; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Pilling, 2010; Baart et al., 2014) or only for the P2 component (Baart et al., 2014). Second, the observed P2 latency facilitation also contrasts with previous studies showing earlier latencies during audio-visual speech perception for both N1 and P2 peaks (van Wassenhove et al., 2005; see also Pilling, 2010, for a small but not consistent effect) or only for N1 peak (Stekelenburg and Vroomen, 2007; Baart et al., 2014; Treille et al., 2014). From these differences, it is hypothesized that N1 and P2 components as well as latency facilitation and amplitude reduction effects might reflect different aspects and/or stages of audio-visual speech integration. For instance, van Wassenhove et al. (2005) observed a visually induced suppression of both N1 and P2 components independently of the visual saliency of the speech stimuli, but a latency reduction of N1 and P2 peaks depending on the degree of their visual predictability. From their results, they argue for two distinct integration stages: (1) a global bimodal perceptual stage, reflected in the amplitude reduction, independent of the featural content of the visual stimulus and possibly reflecting phase-coupling of auditory and visual cortices, and (2) a featural phonetic stage, reflected in the latency facilitation and stronger for P2, in which articulator-specific and predictive visual information

are taking into account in auditory phonetic processing (for further discussion, see van Wassenhove, 2013). In parallel, Stekelenburg and Vroomen (2007), Vroomen and Stekelenburg (2010), and Baart et al. (2014) also argue for a bimodal, non-speech specific stage in audio-visual speech integration but here thought to be reflected in the N1 latency facilitation and amplitude reduction. Congruent with this hypothesis, they observed an amplitude and a latency reduction of auditory-evoked N1 responses during audio-visual perception for both speech and non-speech actions, like clapping hands (Stekelenburg and Vroomen, 2007), as well as for artificial audio-visual stimul, like two moving disks predicting a pure tone when colliding with a fixed rectangle (Vroomen and Stekelenburg, 2010). In addition, they also provided evidence for a P2 amplitude reduction specifically dependent on the phonetic predictability of the visual speech input (Baart et al., 2014; see also Vroomen and Stekelenburg, 2010). Taken together, although the observed differences across the present and previous studies on N1 and/or P2 latency facilitation and/or amplitude reduction are still a matter of debate (van Wassenhove et al., 2005; Baart et al., 2014), they might both reflect multistage processes in audio-visual speech integration and also derive from specific experimental settings used in these studies.

From that latter possibility, one interesting finding is that the observed latency and amplitude reduction in the EEG experiment, notably for the P2 component, did not significantly depend on the degree of visual recognition of the speech targets in the behavioral experiment. This contrasts with two previous studies reporting latency shifts of auditory evoked responses directly function of the visemic information (van Wassenhove et al., 2005; Arnal et al., 2009). For instance, van Wassenhove et al. (2005) demonstrated a visually induced facilitation of the P2 auditory evoked potential which systematically varied according to the visual-only recognition of the presented syllable (i.e., the more visually salient was the syllable, the more stronger the latency facilitation). While they observed a P2 latency facilitation around 25 ms, 16 ms, and 8 ms for /pa/, /ta/, and /ka/ syllables, respectively, we here observed latency facilitations around 17 ms, 13 ms, and 15 ms for the same syllables. However, correlation scores likely depend on overall differences in recognition scores between syllables which were stronger in previous studies (van Wassenhove et al., 2005; Arnal et al., 2009). Furthermore, one important difference between our experimental setting and those used in these two studies is that audio-visual interactions were here tested during live face-to-face interactions between a speaker and a listener, with a unique occurrence of the presented syllable in each trial. This natural stimulus variability contrasts with the limited number of tokens used to represent each syllable in the previous studies which were repeatedly presented to the participants (i.e., van Wassenhove et al. (2005): one speaker, three syllables, one token per syllable and 100 trials per syllable and per modality; Arnal et al. (2009): one speaker, five syllables, one token per syllable and 54 trials per syllable and per modality). Similarly, another possible experimental factor impacting bimodal speech integration comes from the number of syllable type. From that view, it is worthwhile noting that we did observe a latency facilitation during live face-to-face speech perception in our previous study, using a similar experimental design, but only for the N1 component (Treille et al., 2014). In this study, however, a simple two-alternative forced-choice identification task between /pa/ and /ta/ syllables was used. It is therefore possible that specific phonetic contents of these two syllables were less perceptually dominant in this previous study, with a more global yes-no strategy done in relation to the more salient bilabial movements for /pa/ as compared to /ta/ (for experimental designs only using two distinct speech stimuli, see also Stekelenburg and Vroomen, 2007; Pilling, 2010; Vroomen and Stekelenburg, 2010; Baart et al., 2014). Overall, given the significant P2 latency facilitation, our results do not contradict the hypothesis that visual inputs convey predictive information with respect to the incoming auditory speech input (for a discussion on the sensory predictability of audio-visual speech stimuli, see Chandrasekaran et al., 2009; Schwartz and Savariaux, 2013) nor the fact that visual predictability of the speech stimulus might be reflected in auditory evoked responses. We simply argue that visual predictions on the incoming acoustic signal in audio-visual speech perception might likely be constrained not only by the featural content of the visual stimuli but also by the experimental context and by short-term memory traces and knowledge the listener previously acquired on these stimuli.

As in the audio-visual condition, the observed modulation of N1/P2 auditory evoked potentials during audio-haptic speech perception also clearly suggests cross-modal speech interactions between the auditory and the haptic signals. In this bimodal condition, we also observed a latency facilitation on the P2 auditory evoked potential that did not vary according to the degree of haptic recognition of the speech targets. In addition to this latency facilitation, an N1 amplitude reduction was also observed but only for /pa/ syllable. As previously noted, this latter result fits well with a stronger haptic saliency of the bilabial rounding movements involved in /pa/ syllable (see Treille et al., 2014, for behavioral evidence) and with previous studies on audio-visual integration demonstrating that N1 suppression is strongly dependent on whether the visual signal reliably predicts the onset of the auditory event (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010). As discussed previously, the fact that P2 latency reduction was nevertheless observed for all syllables indirectly argue for distinct integration processes in the cortical speech processing hierarchy (van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010; Baart et al., 2014).

Taken together, our results provide new evidence for audio-visual and audio-haptic speech interactions in live dyadic interactions (Treille et al., 2014). The fact that the modulation of N1/P2 auditory evoked potentials were quite similar in these bimodal conditions, despite the less natural haptic modality, further emphasizes the multimodal nature of speech perception. As previously mentioned, apart from speech, multisensory integration from sight, sound and haptic modalities naturally occurs in everyday life. Although bimodal speech perception is a special case of multisensory processing that interfaces with the linguistic system, similar integration processes might have been used to extract temporal and/or phonetic relevant information from the visual and haptic speech signals that, together with the listener's knowledge of speech production (for a review, see Schwartz et al., 2012), might have constrained the incoming auditory processing.

REFERENCES

Alcorn, S. (1932). The Tadoma method. *Volta Rev.* 34, 195–198.

Arnal, L. H., and Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Baart, M., Stekelenburg, J. J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 65, 115–211. doi: 10.1016/j.neuropsychologia.2013.11.011

Benoît, C., Mohamadi, T., and Kandel, S. D. (1994). Effects on phonetic context on audio-visual intelligibility of French. *J. Speech Hear. Res.* 37, 1195–1203.

Besle, J., Fort, A., Delpuech, C., and Giard, M. H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Boersma, P., and Weenink, D. (2013). *Praat: Doing Phonetics by Computer. Computer Program*, Version 5.3.42. Available at: http://www.praat.org/ [accessed March 2, 2013].

Campbell, C. S., and Massaro, D. W. (1997). Perception of visible speech: influence of spatial quantization. *Perception* 26, 627–644. doi: 10.1068/p260627

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Fowler, C., and Dekle, D. (1991). Listening with eye and hand: crossmodal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 816–828. doi: 10.1037/0096-1523.17.3.816

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., and Mühlbauer, M. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *J. Acoust. Soc. Am.* 123, 72–76. doi: 10.1121/1.2884349

Grant, K., Walden, B. E., and Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 103, 2677–2690. doi: 10.1121/1.422788

Green, K. P. (1998). "The use of auditory and visual information during phonetic processing: implications for theories of speech perception," in *Hearing by Eye, II. Perspectives and Directions in Research on Audiovisual Aspects of Language Processing*, eds R. Campbell, B. Dodd, and D. Burnham (Hove: Psychology Press), 3–25.

Hertrich, I., Mathiak, K., Lutzenberger, W., Menning, H., and Ackermann, H. (2007). Sequential audiovisual interactions during speech perception: a whole-head MEG study. *Neuropsychologia* 45, 1342–1354. doi: 10.1016/j.neuropsychologia.2006.09.019

Jones, J. A., and Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Can. Acoust.* 25, 13–19.

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res. Cogn. Brain Res.* 18, 65–75. doi: 10.1016/j.cogbrainres.2003.09.004

Lebib, R., Papo, D., de Bode, S., and Baudonnière, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neurosci. Lett.* 341, 185–188. doi: 10.1016/S0304-3940(03)00131-9

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Näätänen, R., and Picton, T. W. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Navarra, J., and Soto-Faraco, S. (2005). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychol. Res.* 71, 4–12. doi: 10.1007/s00426-005-0031-5

Norton, S. J., Schultz, M. C., Reed, C. M., Braida, L. D., Durlach, N. I., Rabinowitz, W. M., et al. (1977). Analytic study of the Tadoma method: background and preliminary results. *J. Speech Hear. Res.* 20, 574–595.

Pilling, M. (2010). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J. Speech Lang. Hear. Res.* 52, 1073–1081. doi: 10.1044/1092-4388(2009/07-0276)

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard tounderstand: a lipreading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lipreading*, eds R. Campbell and B. Dodd (London: Lawrence Erlbaum Associates), 97–113.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-F

Sato, M., Cavé, C., Ménard, L., and Brasseur, L. (2010). Auditory-tactile speech perception in congenitally blind and sighted adults. *Neuropsychologia* 48, 3683–3686. doi: 10.1016/j.neuropsychologia.2010.08.017

Scherg, M., and Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalogr. Clin. Neurol.* 65, 344–360. doi: 10.1016/0168-5597(86)90014-6

Schwartz, J. L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition* 93, B69–B78. doi: 10.1016/j.cognition.2004.01.006

Schwartz, J. L., Ménard, L., Basirat, A., and Sato, M. (2012). The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. *J. Neurolinguistics* 25, 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Schwartz, J. L., and Savariaux, C. (2013). "Data and simulations about audiovisual asynchrony and predictability in speech perception," in *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing*, Annecy, France.

Stein, B. E. (2012). *The New Handbook of Multisensory Processing*. Cambridge: MIT Press.

Stein, B. E., and Meredith, M. A. (1993). *The New Handbook of Multisensory Processing*. Cambridge, MA: MIT Press.

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.* 19, 1964–1973. doi: 10.1162/jocn.2007.19.12.1964

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Treille, A., Cordeboeuf, C., Vilain, C., and Sato, M. (2014). Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. *Neuropsychologia* 57, 71–77. doi: 10.1016/j.neuropsychologia.2014.02.004

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Front. Psychol.* 4:388. doi: 10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2003). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596. doi: 10.1162/jocn.2009.21308

Winneke, A. H., and Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychol. Aging* 26, 427–438. doi: 10.1037/a0021683

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 March 2014; accepted: 21 April 2014; published online: 13 May 2014.
*Citation: Treille A, Vilain C and Sato M (2014) The sound of your lips: electrophysiological cross-modal interactions during hand-to-face and face-to-face speech perception. Front. Psychol. **5**:420. doi: 10.3389/fpsyg.2014.00420*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

ADVANTAGES OF PUBLISHING IN FRONTIERS

FAST PUBLICATION
Average 90 days from submission to publication

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

RESEARCH NETWORK
Our network increases readership for your article

OPEN ACCESS
Articles are free to read, for greatest visibility

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

GLOBAL SPREAD
Six million monthly page views worldwide

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

IMPACT METRICS
Advanced metrics track your article's impact

SUPPORT
By our Swiss-based editorial team

frontiers

EPFL Innovation Park · Building I · 1015 Lausanne · Switzerland
T +41 21 510 17 00 · info@frontiersin.org · frontiersin.org