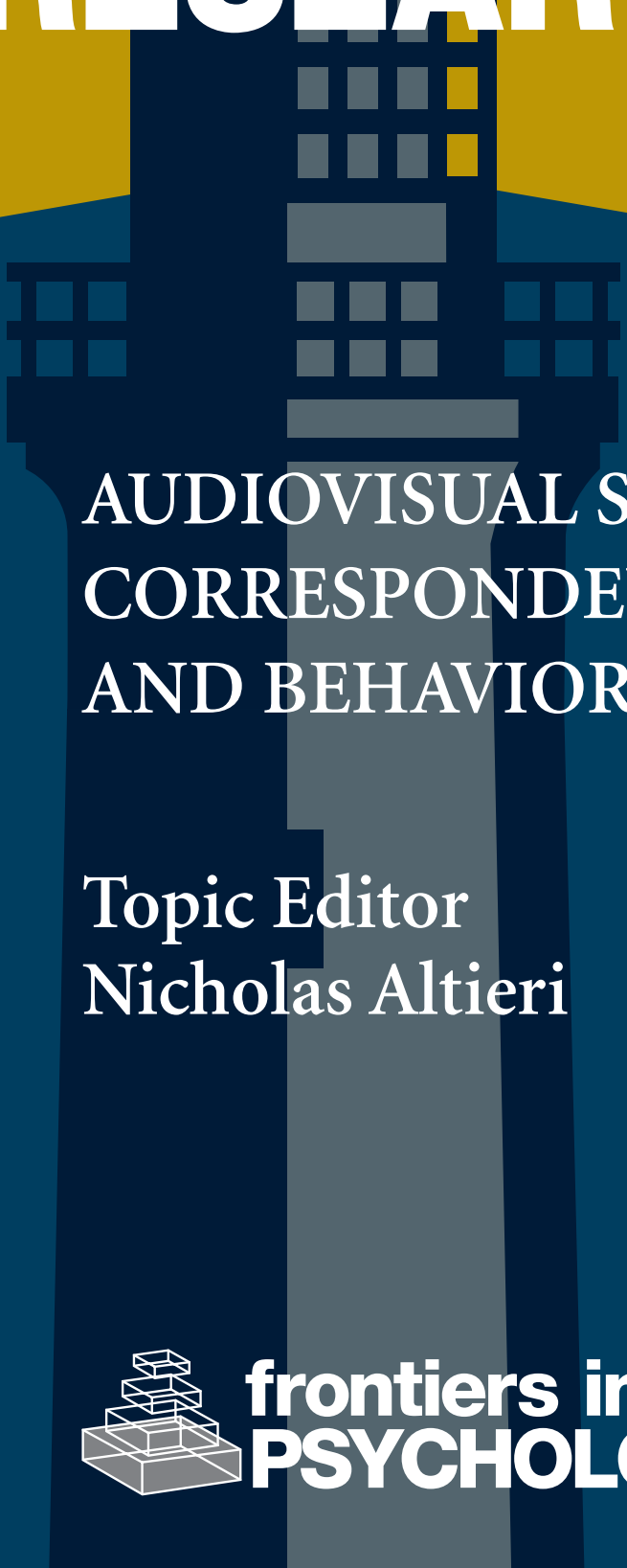
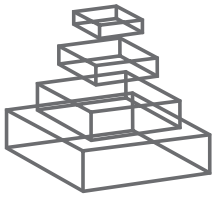
frontiers
RESEARCH TOPICS



AUDIOVISUAL SPEECH RECOGNITION: CORRESPONDENCE BETWEEN BRAIN AND BEHAVIOR

Topic Editor
Nicholas Altieri

frontiers in PSYCHOLOGY



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-251-9
DOI 10.3389/978-2-88919-251-9

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

AUDIOVISUAL SPEECH RECOGNITION: CORRESPONDENCE BETWEEN BRAIN AND BEHAVIOR

Topic Editor:
Nicholas Altieri, Idaho State University, USA

Perceptual processes mediating recognition, including the recognition of objects and spoken words, is inherently multisensory. This is true in spite of the fact that sensory inputs are segregated in early stages of neuro-sensory encoding. In face-to-face communication, for example, auditory information is processed in the cochlea, encoded in auditory sensory nerve, and processed in lower cortical areas. Eventually, these "sounds" are processed in higher cortical pathways such as the auditory cortex where it is perceived as speech. Likewise, visual information obtained from observing a talker's articulators is encoded in lower visual pathways. Subsequently, this information undergoes processing in the visual cortex prior to the extraction of articulatory gestures in higher cortical areas associated with speech and language. As language perception unfolds, information garnered from visual articulators interacts with language processing in multiple brain regions. This occurs via visual projections to auditory, language, and multisensory brain regions. The association of auditory and visual speech signals makes the speech signal a highly "configural" percept.

An important direction for the field is thus to provide ways to measure the extent to which visual speech information influences auditory processing, and likewise, assess how the unisensory components of the signal combine to form a configural/integrated percept. Numerous behavioral measures such as accuracy (e.g., percent correct, susceptibility to the "McGurk Effect") and reaction time (RT) have been employed to assess multisensory integration ability in speech perception. On the other hand, neural based measures such as fMRI, EEG and MEG have been employed to examine the locus and or time-course of integration. The purpose of this Research Topic is to find converging behavioral and neural based assessments of audiovisual integration in speech perception. A further aim is to investigate speech recognition ability in normal hearing, hearing-impaired, and aging populations. As such, the purpose is to obtain neural measures from EEG as well as fMRI that shed light on the neural bases of multisensory processes, while connecting them to model based measures of reaction time and accuracy in the behavioral domain. In doing so, we endeavor to gain a more thorough description of the neural bases and mechanisms underlying integration in higher order processes such as speech and language recognition.

Table of Contents

frontiers in
PSYCHOLOGY

EDITORIAL
published: 17 September 2013
doi: 10.3389/fpsyg.2013.00642

Audiovisual integration: an introduction to behavioral and neuro-cognitive methods

Nicholas Altieri*

Communication Sciences and Disorders, Idaho State University, Pocatello, ID, USA
**Correspondence: altinich@isu.edu*

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

Keywords: audiovisual speech, integration, brain, speech and cognition, neuroimaging of speech, quantitative methods multisensory speech

Advances in neurocognitive and quantitative behavioral techniques have offered new insights to the study of cognition and language perception. This includes ways in which neurological processes and behavior are intimately intertwined. Examining traditional behavioral measures and model predictions, along with neurocognitive measures, will provide a powerful theory-driven and unified approach for researchers in the cognitive and language sciences. In this topic, the aim was to highlight some of the noteworthy methodological developments in the burgeoning field of multisensory speech perception.

Decades of research on audiovisual speech integration has, broadly speaking, reshaped the way language processing is conceptualized in the field. Beginning with Sumby and Pollack's seminal study of audiovisual integration published in 1954, qualitative and quantitative relationships have emerged showing the benefit of being able to obtain visual cues from "speech reading" under noisy conditions. A pioneering study by McGurk and MacDonald (1976) further demonstrated a form of integration phenomenon in which incongruent auditory-visual speech signals contribute to a fused or combined percept. (One such example is an auditory "ba" dubbed over a video of a talker articulating the syllable "ga." This often yields a combined percept of "da.")

Methods for determining whether "integration" occurs have, for example, involved examining whether a listener is susceptible to the McGurk effect, as we shall in a study by Setti et al. (2013) in the Research Topic. Perhaps a more commonly used assessment tool for determining the presence of "integration" has been measuring the extent to which a dependent variable (accuracy, speed, etc.) obtained from audiovisual trials is significantly "better" than the predicted response obtained from the unisensory conditions. A difference between obtained and predicted measures is thought to indicate a violation of independence between modalities (Altieri and Townsend, 2011; Altieri et al., 2013). In recent years, the neurological bases of these multisensory phenomena in speech perception have been developed largely in parallel with advances in behavioral techniques. Neuroimaging studies have looked at the Blood Oxygen-Level Dependent (BOLD) signal in relation to AV speech stimuli and compared that to the unisensory BOLD responses (e.g., Calvert, 2001; Stevenson and James, 2009). Within the milieu of EEG studies, similar comparisons have been made between the amplitude evoked by audiovisual, vs. auditory and visual-only stimuli. Similar to the fMRI studies, EEG research has contributed to the idea that integration occurs if the AV response differs from the unisensory responses ($AV_{ERP} < A_{ERP} + V_{ERP}$; see, van Wassenhove et al., 2005; and Winneke and Phillips, 2011).

The application of EEG, fMRI or other imaging techniques in combination with behavioral indexes has therefore enhanced the testability of neural based theories of multisensory language processing. The broader aim of this Research Topic was to investigate the variety of manners in which neural measures of multisensory language processing could be anchored to behavioral indices of integration.

Several pioneering studies appear in this volume addressing a wide variety of issues in multisensory speech recognition. Quite significantly, this research explores integration in different age groups, for individuals with sensory processing deficits, and across different listening environments. First, a study carried out by Altieri and Wenger (2013) sought to rigorously associate the dynamic psychophysical measures of perception—namely the reaction time measure of *workload capacity* (Townsend and Nozawa, 1995)—with a neural dynamics from EEG. Under degraded listening conditions, we observed an increase in integration efficiency as measured by capacity, which co-occurred with an increase in multisensory ERPs relative to auditory-only ERPs. In a much needed review on the rules giving rise to multisensory integration, van Wassenhove (2013) provided an overview of "predictive coding hypotheses." Updated hypotheses were considered, namely concerning how internal predictions about linguistics percepts are formulated. An overview of neuroimaging literature was included in the discussion.

Three reports explored the temporal effects of visual information on auditory encoding. One, provided by Ten Oever et al. (2013), varied the synchrony of the auditory and visual signals to explore the temporal effects of auditory syllable encoding. The results indicated a larger time-window for congruent AV syllables. Second, Moradi et al. (2013) provided a report investigating the influence of visual information on temporal recognition. This study showed that visual cues sped-up linguistic recognition in both noisy and clear listening conditions. Finally, a review and hypothesis article by Hertrich et al. (2013) proposes a brain network explaining how blind individuals, on average, are capable of perceiving auditory speech at a much faster rate compared to individuals with normal vision. Together, these articles will help constrain dynamic and neural-based theories regarding temporal aspects of audiovisual speech perception.

Two studies in this Research Topic also explored the effects of aging and neural development on perceptual skills. Kushnerenko et al. (2013) used an eye tracking paradigm in conjunction with

ERPs to investigate the extent to which these measures predict normal linguistic development in children. Second, Setti et al. (2013) investigated integration skills by looking at whether age is predictive of the susceptibility to the McGurk effect. Interestingly, the authors found that older adults were more susceptible to the fusion than younger ones—ostensibly due to differences in perceptual rather than higher order cognitive processing abilities.

These research and review articles provide a rich introduction to a variety of fascinating techniques for investigating speech integration. Ideally, these research directions will pave the way toward a much improved tapestry of methodologies, and refinements of neuro-cognitive theories of multisensory processing across life-span, listening conditions, and sensory-cognitive abilities.

REFERENCES

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. (2013). A dynamic assessment function for measuring age-related sensory decline in audiovisual speech recognition. *Behav. Res. Methods.* doi: 10.3758/s13428-013-0372-8. [Epub ahead of print].

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

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

Hertrich, I., Dietrich, S., and Ackermann, H. (2013). How can audiovisual pathways enhance the temporal resolution of time-compressed speech in blind subjects. *Front. Psychol.* 4:530. doi: 10.3389/fpsyg.2013.00530

Kushnerenko, E. V., Tomalski, P., Ballieux, H., Potton, A., Birtles, D., Frostick, C., et al. (2013). Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life. *Front. Psychol.* 4:432. doi: 10.3389/fpsyg.2013.00432

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

Moradi, S., Lidestam, B., and Rönnberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. *Front. Psychol.* 4:359. doi: 10.3389/fpsyg.2013.00359

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

Stevenson, R. A., and James, T. W. (2009). Neuronal convergence and inverse effectiveness with audiovisual integration of speech and tools in human superior temporal sulcus: evidence from BOLD fMRI. *Neuroimage* 44, 1210–1223. doi: 10.1016/j.neuroimage.2008.09.034

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

Ten Oever, S., Sack, A. T., 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

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

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., 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

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

Received: 23 August 2013; accepted: 29 August 2013; published online: 17 September 2013.

*Citation: Altieri N (2013) Audiovisual integration: an introduction to behavioral and neuro-cognitive methods. Front. Psychol. **4**:642. doi: 10.3389/fpsyg.2013.00642*

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

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 12 July 2013
doi: 10.3389/fpsyg.2013.00388

Speech through ears and eyes: interfacing the senses with the supramodal brain

Virginie van Wassenhove [1,2,3]*

[1] Cognitive Neuroimaging Unit, Brain Dynamics, INSERM, U992, Gif/Yvette, France
[2] NeuroSpin Center, CEA,DSV/I2BM, Gif/Yvette, France
[3] Cognitive Neuroimaging Unit, University Paris-Sud, Gif/Yvette, France

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
Nicholas Altieri, Idaho State University, USA
Luc H. Arnal, New York University, USA

****Correspondence:***
Virginie van Wassenhove, CEA/DSV/I2BM/Neurospin, Bât 145 Point courrier 156, Gif/Yvette 91191, France
e-mail: Virginie.van-Wassenhove@cea.fr

The comprehension of auditory-visual (AV) speech integration has greatly benefited from recent advances in neurosciences and multisensory research. AV speech integration raises numerous questions relevant to the computational rules needed for binding information (within and across sensory modalities), the representational format in which speech information is encoded in the brain (e.g., auditory vs. articulatory), or how AV speech ultimately interfaces with the linguistic system. The following non-exhaustive review provides a set of empirical findings and theoretical questions that have fed the original proposal for predictive coding in AV speech processing. More recently, predictive coding has pervaded many fields of inquiries and positively reinforced the need to refine the notion of internal models in the brain together with their implications for the interpretation of neural activity recorded with various neuroimaging techniques. However, it is argued here that the strength of predictive coding frameworks reside in the specificity of the generative internal models not in their generality; specifically, internal models come with a set of rules applied on particular representational formats themselves depending on the levels and the network structure at which predictive operations occur. As such, predictive coding in AV speech owes to specify the level(s) and the kinds of internal predictions that are necessary to account for the perceptual benefits or illusions observed in the field. Among those specifications, the actual content of a prediction comes first and foremost, followed by the representational granularity of that prediction in time. This review specifically presents a focused discussion on these issues.

Keywords: analysis-by-synthesis, predictive coding, multisensory integration, Bayesian priors

INTRODUCTION

In natural conversational settings, watching an interlocutor's face does not solely provide information about the speaker's identity or emotional state: the kinematics of the face articulating speech can robustly influence the processing and comprehension of auditory speech. Although audiovisual (AV) speech perception is ecologically relevant, classic models of speech processing have predominantly accounted for speech processing on the basis of acoustic inputs (e.g., **Figure 1**). From an evolutionary standpoint, proximal communication naturally engages multisensory interactions i.e., vision, audition, and touch but it is not until recently that multisensory integration in the communication system of primates has started to be investigated neurophysiologically (Ghazanfar and Logothetis, 2003; Barraclough et al., 2005; Ghazanfar et al., 2005, 2008; Kayser et al., 2007, 2010; Kayser and Logothetis, 2009; Arnal and Giraud, 2012). Advances in multisensory research has raised core issues: how early do multisensory integration occur during perceptual processing (Talsma et al., 2010)? In which representational format do sensory modalities interface for supramodal (Pascual-Leone and Hamilton, 2001; Voss and Zatorre, 2012) and speech analysis (Summerfield, 1987; Altieri et al., 2011)? Which neuroanatomical pathways are implicated (Calvert and Thesen, 2004; Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008; Murray and Spierer, 2011)? In Humans, visual speech plays an important role in social interactions (de Gelder et al., 1999) but also, and crucially, interfaces with the language system at various depth of linguistic processing (e.g., McGurk and MacDonald, 1976; Auer, 2002; Brancazio, 2004; Campbell, 2008). AV speech thus provides an appropriate model to address the emergence of supramodal or abstract representations in the Human mind and to build upon a rich theoretical and empirical framework elaborated in linguistic research in general (Chomsky, 2000) and in speech research, in particular (Chomsky and Halle, 1968; Liberman and Mattingly, 1985).

WEIGHTING SENSORY EVIDENCE AGAINST INTERNAL NON-INVARIANCE

Speech theories have seldom incorporated visual information as raw material for speech processing (Green, 1996; Schwartz et al., 1998) although normal hearing and hearing-impaired populations greatly benefit from looking at the interlocutor's face (Sumby and Pollack, 1954; Erber, 1978; MacLeod and Summerfield, 1987; Grant and Seitz, 1998, 2000). If any benefit

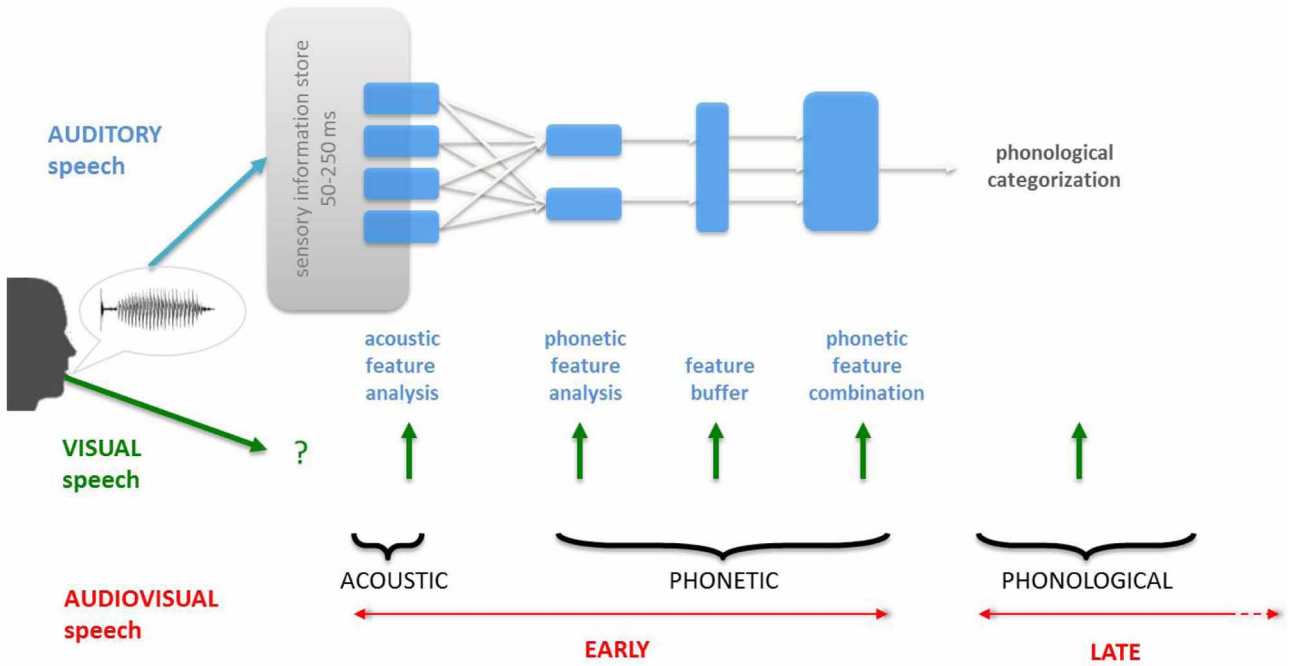


FIGURE 1 | Classic information-theoretic description of speech processing. Classic models of speech processing have been construed on the basis of the acoustics of speech, leaving aside the important contribution of visual speech inputs. As a result, the main question in audiovisual (AV) speech processing has been: when does visual speech information integrate with auditory speech? The two main alternatives are before (acoustic or phonetic features, "early" integration) or after ("late" integration) the phonological categorization of the auditory speech inputs (see also Schwartz et al., 1998). However, this model unrealistically frames and biases the question of "when" by imposing a serial, linear and hierarchical processing for speech processing.

for speech encoding is to be gained in the integration of AV information, the informational content provided by each sensory modality is likely to be partially, but not solely, redundant i.e., complementary. For instance, the efficiency in AV speech integration is known to depend not only on the amount of information extracted in each sensory modality but also in its variability (Grant et al., 1998). Understanding the limitations and processing constraints of each sensory modality is thus important to understand how non-invariance in speech signals leads to invariant representations in the brain. In that regards, should speech processing be considered "special?" The historical debate is outside the scope of this review but it is here considered that positing an internal model dedicated to the processing of speech analysis is legitimate to account for (i) the need for invariant representations in the brain, (ii) the parsimonious sharing of generative rules for perception/production and (iii) the ultimate interfacing of the (AV) communication system with the Human linguistic system. As such, this review focuses on the specificities of AV speech not on the general guiding principles of multisensory (AV) integration.

TEMPORAL PARSING AND NON-INVARIANCE

A canonical puzzle in (auditory, visual and AV) speech processing is how the brain correctly parses a continuous flow of sensory information. Like auditory speech, the visible kinematics of articulatory gestures hardly provides non-invariant structuring of information over time (Kent, 1983; Tuller and Kelso, 1984; Saltzman and Munhall, 1989; Schwartz et al., 2012) yet temporal information in speech is critical (Rosen, 1992; Greenberg, 1998). Auditory speech is typically sufficient to provide a high level of intelligibility (e.g., over the phone) and accordingly, the auditory system can parse incoming speech information with high-temporal acuity (Poeppel, 2003; Morillon et al., 2010; Giraud and Poeppel, 2012). Conversely, visual speech alone leads to poor intelligibility scores (Campbell, 1989; Massaro, 1998) and visual processing is characterized by a slower sampling rate (Busch and VanRullen, 2010). The slow timescales over which visible articulatory gestures evolve (and are extracted by the observer's brain) constrain the representational granularity of visual information to visemes, categories much less distinctive than phonemes.

In auditory neuroscience, the specificity of phonetic processing and phonological categorization has long been investigated (Maiste et al., 1995; Simos et al., 1998; Liégeois et al., 1999; Sharma and Dorman, 1999; Philips et al., 2000). The peripheral mammalian auditory system has been proposed to efficiently encode a broad category of natural acoustic signals by using a time-frequency representation (Lewicki, 2002; Smith and Lewicki, 2006). In this body of work, the characteristics of auditory filters heavily depend on the statistical characteristics of sounds: as such, auditory neural coding schemes show plasticity as a function of acoustic inputs. The intrinsic neural tuning properties allow for multiple modes of acoustic processing with trade-offs in the time and frequency domains

which naturally partition the time-frequency space into sub-regions. Complementary findings show that efficient coding can be realized for speech inputs (Smith and Lewicki, 2006) supporting the notion that the statistical properties of auditory speech can drive different modes of information extraction in the same neural populations, an observation supporting the "speech mode" hypothesis (Remez et al., 1998; Tuomainen et al., 2005; Stekelenburg and Vroomen, 2012).

In visual speech, how the brain derives speech-relevant information from seeing the dynamics of the facial articulators remains unclear. While the neuropsychology of lipreading has been thoroughly described (Campbell, 1986, 1989, 1992), very few studies have specifically addressed the neural underpinnings of visual speech processing (Calvert, 1997; Calvert and Campbell, 2003). Visual speech is a particular form of biological motion which readily engages some face-specific sub-processes (Campbell, 1986, 1992) but remains functionally independent from typical face processing modules (Campbell, 1992). Insights on the neural bases of visual speech processing may be provided by studies of biological motion (Grossman et al., 2000; Vaina et al., 2001; Servos et al., 2002) and the finding of mouth-movement specific cells in temporal cortex provides a complementary departing point (Desimone and Gross, 1979; Puce et al., 1998; Hans-Otto, 2001). Additionally, case studies (sp. prosopagnosia and akinetopsia) have suggested that both form and motion are necessary for the processing of visual and AV speech (Campbell et al., 1990; Campbell, 1992). In line with this, an unexplored hypothesis for the neural encoding of facial kinematics is the use form-from-motion computations (Cathiard and Abry, 2007) which could help the implicit recovery of articulatory commands from seeing the speaking face (e.g., Viviani et al., 2011).

ACTIVE SAMPLING OF VISUAL SPEECH CUES

In spite of the limited informational content provided by visual speech (most articulatory gestures remain hidden), AV speech integration is resilient to further degradation of the visual speech signal. Numerous filtering approaches do not suppress integration (Rosenblum and Saldaña, 1996; Campbell and Massaro, 1997; Jordan et al., 2000; MacDonald et al., 2000) suggesting that the use of multiple visual cues [e.g., luminance patterns (Jordan et al., 2000); kinematics (Rosenblum and Saldaña, 1996)]. Additionally, neither the gender (Walker et al., 1995) nor the familiarity (Rosenblum and Yakel, 2001) of the face impacts the robustness of AV speech integration. As will be discussed later, AV speech integration also remains resilient to large AV asynchronies (cf. *Resilient temporal integration and the co-modulation hypothesis*). Visual kinematics alone are sufficient to maintain a high rate of AV integration (Rosenblum and Saldaña, 1996) but whether foveal (i.e., explicit lip-reading with focus on the mouth area) or extra-foveal (e.g., global kinematics) information is most relevant for visemic categorization remains unclear.

Interestingly, gaze fixations 10–20° away from the mouth are sufficient to extract relevant speech information but numerous eye movements have also been reported (Vatikiotis-Bateson et al., 1998; Paré et al., 2003). It is noteworthy that changes of gaze direction can be crucial for the extraction of auditory information as neural tuning properties throughout the auditory pathway are modulated by gaze direction (Werner-Reiss et al., 2003) and auditory responses are affected by changes in visual fixations (Rajkai et al., 2008; van Wassenhove et al., 2012). These results suggest an interesting working hypothesis: the active scanning of a speaker's face may compensate for the slow sampling rate of the visual system.

Hence, despite the impoverished signals provided by visual speech, additional degradation does not fully prevent AV speech integration. As such, (supramodal) AV speech processing is more likely than not a natural mode of processing in which the contribution of visual speech to the perceptual outcome may be regulated as a function of the needs for perceptual completion in the system.

AV SPEECH MODE HYPOTHESIS

Several findings have suggested that AV signals displayed in a speech vs. a non-speech mode influence both behavioral and electrophysiological responses (Tuomainen et al., 2005; Stekelenburg and Vroomen, 2012). Several observations could complement this view. First, lip-reading stands as a natural ability that is difficult to improve (as opposed to reading ability; Campbell, 1992) and is a good predictor of AV speech integration (Grant et al., 1998). In line with these observations, and as will be discussed later on, AV speech integration undergoes a critical acquisition period (Schorr et al., 2005).

Second, within the context of an internal speech model, AV speech integration is not arbitrary and follows principled internal rules. In the seminal work of McGurk and MacDonald (1976, MacDonald and McGurk, 1978), two types of phenomena illustrate principled ways in which AV speech integration occurs. In *fusion*, dubbing an auditory bilabial (e.g., [ba] or [pa]) onto a visual velar place of articulation (e.g., [ga] or [ka]) leads to an illusory fused alveolar percept (e.g., [da] or [ta], respectively). Conversely, in *combination*, dubbing an auditory [ga] onto a visual place of articulation [ba] leads to the illusory combination percept [bga]. Fusion has been used as an index of automatic AV speech integration because it leads to a unique perceptual outcome that is nothing like any of the original sensory inputs (i.e., neither a [ga] nor a [ba], but a third percept). Combination has been much less studied: unlike fusion, the resulting percept is not unique but rather a product of co-articulated speech information (such as [bga]). Both fusion and combination provide convenient (albeit arguable) indices on whether AV speech integration has occurred or not. These effects can be generalized across places-of-articulation in stop-consonants such that any auditory bilabial dubbed onto a visual velar result in a misperceived alveolar. These two kinds of illusory AV speech outputs illustrate the complexity of AV interactions and suggest that the informational content carried by each sensory modality determines the nature of AV interactions during speech processing. A strong hypothesis is that internal principles should depend on the articulatory repertoire of a given language and few cross-linguistic studies have addressed this issue (Sekiyama and Tohkura, 1991; Sekiyama, 1994, 1997).

Inherent to the speech mode hypothesis is the attentional-independence of speech analysis. Automaticity in AV speech processing (and in multisensory integration) is a matter of great

debate (Talsma et al., 2010). A recent finding (Alsius and Munhall, 2013) suggests that conscious awareness of a face is not necessary for McGurk effects (cf. also Vidal et al. submitted, pers. communication). While attention may regulate the weight of sensory information being processed in each sensory modality—e.g., via selective attention (Lakatos et al., 2008; Schroeder and Lakatos, 2009)—attention does not a priori overtake the internal generative rules for speech processing. In other words, while the strength of AV speech integration can be modulated (Tiippana et al., 2003; Soto-Faraco et al., 2004; Alsius et al., 2005; van Wassenhove et al., 2005), AV speech integration is not fully abolished in integrators.

The robustness and principled ways in which visual speech influences auditory speech processing suggest that the neural underpinnings of AV speech integration rely on specific computational mechanisms that are constrained by the internal rules of the speech processing system—and possibly modulated by attentional focus on one or the other streams of information. I now elaborate on possible predictive implementations and tenants of AV speech integration.

PREDICTIVE CODING, PRIORS AND THE BAYESIAN BRAIN

A majority of mental operations are cognitively impenetrable i.e., inaccessible to conscious awareness (Pylyshyn, 1984; Kihlstrom, 1987). Proposed more than a century ago [Parrot (cf. Allik and Konstabel, 2005); Helmholtz MacKay, 1958; Barlow, 1990; Wundt (1874)], unconscious inferences later coined the role of sensory processing as a means to remove redundant information in the incoming signals based on the informed natural statistics of sensory events. For instance, efficient coding disambiguates incoming sensory information using mutual inhibition as a means to decorrelate mixed signals: a network can locally generate hypotheses on the basis of a known (learned) matrix from which inversion can be drawn for prediction (Barlow, 1961; Srinivasan et al., 1982; Barlow and Földiak, 1989). Predictive coding can be local, for instance with a specific instantiation in the architecture of the retina (Hosoya et al., 2005). Early predictive models have essentially focused on the removal of redundant information in the spatial domain. Recently, predictive models have incorporated more sophisticated levels of predictions (Harth et al., 1987; Rao and Ballard, 1999; Friston, 2005). For instance, Harth et al. (1987) proposed a predictive model in which feedback connectivity shapes the extraction of information early in the visual hierarchy and such regulation of V1 activity in the analysis of sensory inputs has also been tested (Sharma et al., 2003). The initial conception of "top–down" regulation has been complemented with the notion that feedforward connections may not carry the extracted information *per se* but rather the residual error between "top–down" internal predictions and the incoming sensory evidence (Rao and Ballard, 1999).

A growing body of evidence supports the view that the brain is a hierarchically organized inferential system in which internal hypotheses or predictions are generated at higher levels and tested against evidence at lower levels along the neural pathways (Friston, 2005): predictions are carried by backward and lateral connections whereas prediction errors are carried by forward projections. Predictive coding schemes have thus gone from local circuitries to brain system seemingly suggesting that access to high-level representations are necessary to formulate efficient predictions.

FIXED vs. INFORMED PRIORS

Conservatively, any architectural constraint (e.g., connectivity pattern, gross neuroanatomical pathways), knowledge and circuitry acquired during a sensitive and before a critical period, or the endowment of the system can all be considered deterministic or *fixed priors*. Contrariwise, *informed priors* are any form of knowledge undergoing updates available through plastic changes and acquired through experience.

At the system level, a common neurophysiological index taken as evidence for predictive coding in cortex is the MisMatch Negativity (MMN) response (Näätänen et al., 1978; Näätänen, 1995): the MMN is classically elicited by the presentation of a rare event (∼20% of the time) in the context of standard events (∼80% of the time). The most convincing evidence for the MMN as a residual error resulting from the comparison of an internal prediction with incoming sensory evidence is the case of the MMN to omission, namely an MMN elicited when an event is omitted in a predictable sequence of events (Tervaniemi et al., 1994; Yabe et al., 1997; Czigler et al., 2006). Other classes of electrophysiological responses have been interpreted as residual errors elicited by a deviance at different levels of perceptual or linguistic complexities (e.g., the N400; Lau et al., 2008). Recent findings have also pointed out to the hierarchical level at which statistical contingencies can be incorporated in a predictive model (Wacongne et al., 2011). Altogether, these results are in line with recent hierarchical processing of predictive coding in which the complexity of the prediction depends on the depth of recursion in the predictive model (Kiebel et al., 2008).

In AV speech, the seminal work of Sams and Aulanko (1991) used an MMN paradigm with magnetoencephalography (MEG). Using congruent and incongruent (McGurk: audio [pa] dubbed onto visual [ka]) stimuli, the authors found that the presentation of an incongruent (congruent) AV speech deviant in a stream of congruent (incongruent) AV speech standards elicited a robust auditory MMN. Since, a series of subsequent MMN studies has replicated these findings (Colin et al., 2002; Möttönen et al., 2002, 2004) and the sources of the MMN was consistently located in auditory association areas, about 150 to 200 ms following auditory onset and in the superior temporal sulcus from 250 ms on. The bulk of literature using MMN in AV speech therefore suggests that internal predictions generated in the auditory regions incorporate visual information relevant for the analysis of speech.

Critically, it is here argued that internal models invoked for speech processing are part of the cognitive architecture i.e., likely endowed with fixed priors for the analysis of (speech) inputs. The benefit of positing an internal model is precisely to account for robust and invariant internal representations that are resilient to the ever-changing fluctuations of a sensory environment. As such, a predictive model should help refine the internal representations in light of sensory evidence, not entirely shape the internal prediction on the basis of the temporary environmental statistics.

In this context, the temporal statistics of stimuli using an MMN paradigm (e.g., 80% standards, 20% deviants) confine predictions to the temporary experimental context: the residual error is context-specific and tied to the temporary statistics of inputs provided within a particular experimental session. Thus, the MMN may not necessarily reveal fixed priors or specific hard-wired constrains of the system. An internal model should provide a means to stabilize non-invariance in order to counteract the highly variable nature of speech utterances irrespective of the temporally local context. A strong prediction is thus that the fixed priors of an internal model should supersede the temporary statistics of stimuli during a particular experimental session. Specifically, if predictive coding is a canonical operation of cortical function, residual errors should be the rule, not the exception and residual errors should be informative with respect to the content of the prediction, not only with respect to the temporal statistics of the sensory evidence. Following this observation, an experimental design using an equal number of different types of stimuli should reveal predictive coding indices that specifically target the hard-constraints or fixed priors of the system. In AV speech, auditory event-related potentials elicited by the presentation of AV speech stimuli show dependencies on the content of visual speech stimuli: auditory event-related potentials could thus be interpreted as the resulting residual-errors of a comparison process between auditory and visual speech inputs (van Wassenhove et al., 2005).

The argument elaborated here is that to enable a clear interpretation of neurophysiological and neuroimaging data using predictive approaches, the description of the internal model being tested along with the levels at which predictions are expected to occur (hence, the representational format and content of the internal predictors) has become necessary. For instance, previous electrophysiological indices of AV speech integration (van Wassenhove et al., 2005) including latency (interpreted as visual modulations of auditory responses that are speech content-dependent) and amplitude (interpreted as visual modulations of auditory responses that are speech content-independent) effects are not incompatible with the amplitude effects reported in other studies (e.g., Stekelenburg and Vroomen, 2007). AV speech integration implicates speech-specific predictions (e.g., phonetic, syllabic, articulatory representations) but also entails more general operations such as temporal expectation or attentional modulation. As such, the latency effects showed speech selectivity whereas amplitude effects did not; the former may index speech-content predictions coupled with temporal expectations, whereas the latter may inform on general predictive rules. Hierarchical levels can operate predictively in a non-exclusive and parallel manner. The benefit of predictive coding approaches is thus the refinement internal generative models, their specificity with regards to the combinatorial rules that are being used and the representational formats and contents of the different levels of predictions implicated in the model.

BAYESIAN IMPLEMENTATION OF PREDICTIVE CODING

Can Bayesian computations serve predictive coding for speech processing? Recent advances in computational neurosciences have offered a wealth of insights on the Bayesian brain (Denève and Pouget, 2004; Ernst and Bülthoff, 2004; Ma et al., 2006; Yuille and Kersten, 2006) and have opened new and essential venues for the interpretation of perceptual and cognitive operations.

AV speech research has seen the emergence of one of the first Bayesian models for perception, the Fuzzy Logical Model of Perception or FLMP (Massaro, 1987, 1998). In the initial FLMP, the detection and the evaluation stages in speech processing were independent and eventually merged into a single evaluation process (Massaro, 1998). At this level, each speech signals is independently evaluated against prototypes in memory store and assigned a "fuzzy truth value" representing how well the input matches a given prototype. The fuzzy truth value could range from 0 (does not match at all) to 1 (exactly matches the prototype); the prototypical feature represents the ideal value that an exemplar of the prototype holds—i.e., 1 in fuzzy logic—hence the probability that a feature is present in the speech inputs. The prototypes are defined as speech categories which provide an ensemble of features and their conjunctions (Massaro, 1987). In AV speech processing, the 0 to 1 mapping in each sensory modality allowed the use of Bayesian conditional probabilities and computations would take the following form: what is the probability that an AV speech input is a [ba] given a 0.6 probability of being a bilabial in the auditory domain and a 0.7 probability in the visual domain? The best outcome is selected based on the goodness-of-fit determined by prior evidence through a maximum likelihood procedure. Hence, in this scheme, the independence of sensory modalities is necessary to allow the combination of two feature estimates (e.g., place-of-articulations) and a compromise is reached at the decision stage through adjustments of the model with additional sensory evidence. In the FLMP, phonological categorization is thus replaced by a syllabic-like stage (and word structuring) as constrained by the classic phonological rules.

A major criticism of this early Bayesian model for speech perception pertains to the fitting adjustments of the FLMP which would either overfit or be inappropriate for the purpose of predicting integration (Grant, 2002; Schwartz, 2003). Additional discussions have pointed out to the lack of clear accounting of the format of auditory and visual speech representations in such models (Altieri et al., 2011). More recent proposals have notably proposed a parallel architecture to account for AV speech integration efficiency in line with the interplay of inhibitory and excitatory effects seen in neuroimaging data (Altieri and Townsend, 2011).

ANALYSIS-BY-SYNTHESIS (ABYS)

In the seminal description of Analysis-by-Synthesis (AbyS, **Figure 2**) for auditory speech processing by Halle and Stevens (1962), and in line with the Motor Theory of Speech Perception (Liberman et al., 1967; Liberman and Mattingly, 1985), the internal representations used for the production and perception of speech are shared. Specifically, AbyS sketched a predictive implementation for the analysis of auditory speech: the internalized rules for speech production enable to generate hypotheses about which acoustic inputs would come next (Stevens, 1960). From a computational standpoint, AbyS provides the representational system and the fixed priors (internal rules) constraining the

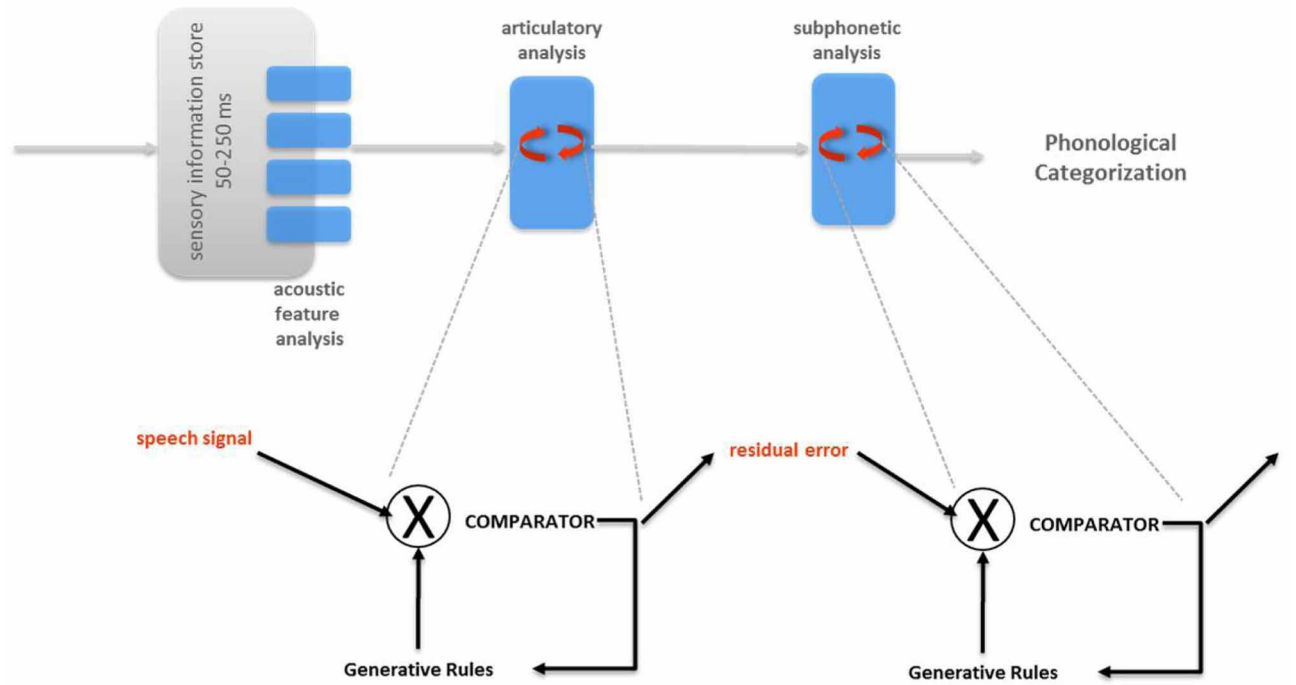


FIGURE 2 | Analysis-by-synthesis (Halle and Stevens, 1962). In the original proposal, two major successive predictive modules are postulated: articulatory analysis followed by subphonetic analysis. In both modules, the generative rules of speech production are used to emit and refine predictions of the incoming sensory signal (articulatory analysis) or residual error from the previous stage (subphonetic analysis).

computations of Bayesian probabilities at the comparison stages. The comparison of auditory and visual speech inputs with internalized articulatory commands can be compatible with Bayesian computations.

In the AbyS, auditory inputs (after preliminary spectral analysis Poeppel et al., 2008) are matched against the internal articulatory rules that would be used to produce the utterance (Halle and Stevens, 1962). Internal speech production rules can take upon continuous values as the set of commands in speech production change as a function of time but "a given articulatory configuration may not be reached before the motion toward the next must be initiated" (Halle and Stevens, 1962). Although the internal rules provide a continuous evaluation of the parameters, the evaluation process can operate on a different temporal scale thereby the units of speech remain discrete and articulatory based. By analogy with the overlap of articulatory commands, the auditory speech inputs contain the traces of preceding and following context (namely, co-articulation effects). Hence, the continuous assignment of values need not bear a one-to-one relationship with the original input signals and overlapping streams of information extraction (for instance, via temporal encoding windows) may enable this process.

AMODAL PREDICTIONS

This early model provided one possible implementation for a forward in time and predictive view of sensory analysis (Stevens, 1960; Halle and Stevens, 1962). Since, AbyS has been re-evaluated in light of recent evidence for predictive coding in speech perception (Poeppel et al., 2008). The internally generated hypotheses are constrained by phonological rules and their distinctive features serve as the discrete units for speech production/perception (Poeppel et al., 2008). The non-invariance of incoming speech inputs can be compensated for by the existence of trading cues matched against the invariant built-in internal rules of the speech system. In particular, the outcome of the comparison process (i.e., the residual error) enables an active correction of the perceptual outcome (i.e., recalibrating so as to match the best fitting value) of the production output.

In conversational settings, the visible articulatory gestures for speech production have recently been argued to precede the auditory utterance by an average of 100–300 ms (Chandrasekaran et al., 2009). The natural precedence of visual speech features could initiate the generation of internal hypotheses as to the incoming auditory speech inputs. This working hypothesis was tested with EEG and MEG by comparing the auditory evoked-responses elicited by auditory and AV speech stimuli (van Wassenhove et al., 2005; **Figure 3**). The early auditory evoked responses elicited by AV speech showed (i) shorter latencies and (ii) reduced amplitudes compared to those elicited by auditory speech alone (van Wassenhove et al., 2005; Arnal et al., 2009). Crucially, the latency shortening of auditory evoked responses was a function of the ease with which participants categorized visual speech alone, thereby a [pa] lead to shorter latencies than [ka] or [ta]. In the context of AbyS, the reliability with which visual speech can trigger internal predictions for incoming auditory speech constrains the analysis of auditory speech (van

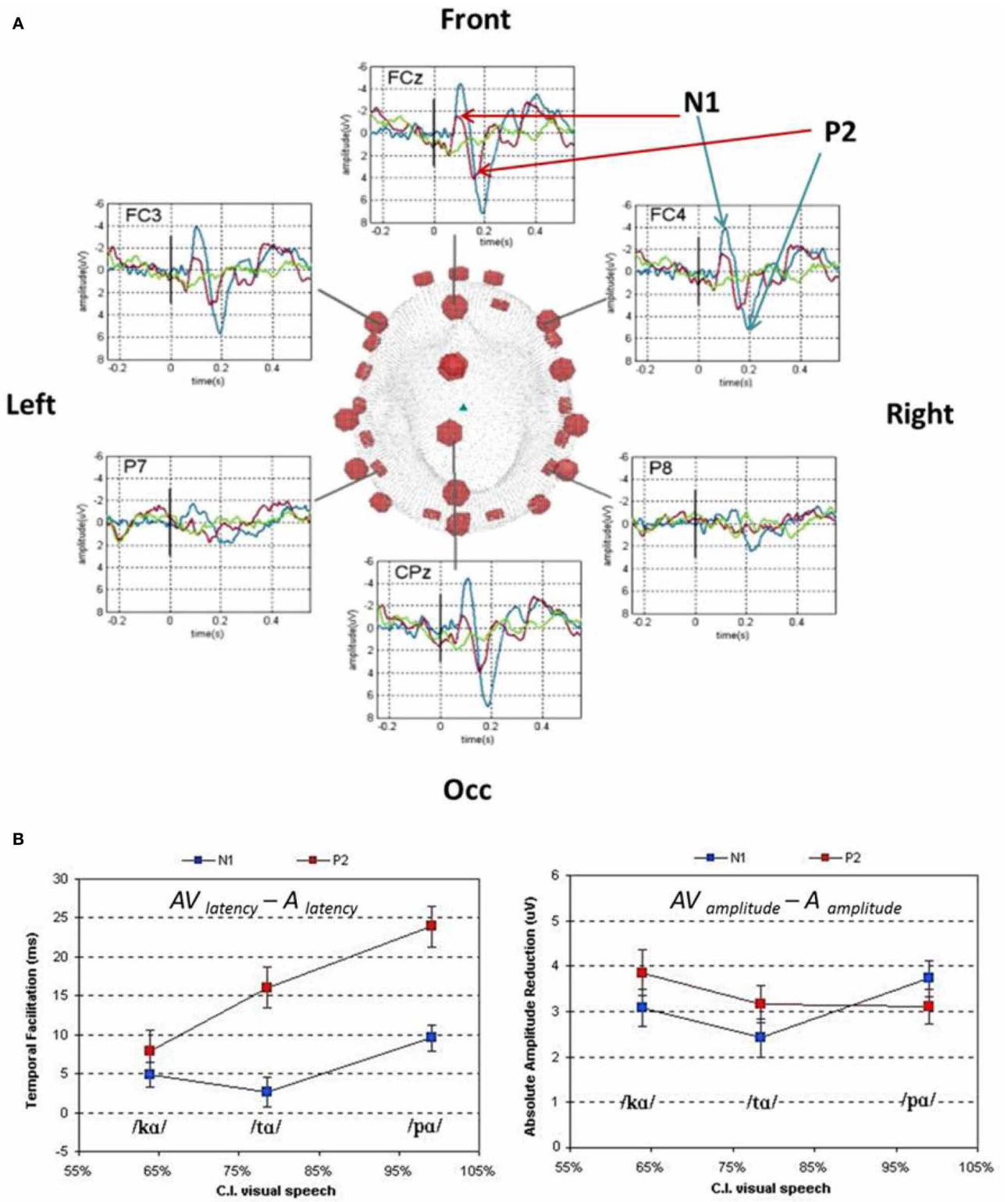


FIGURE 3 | Auditory event-related potentials in response to auditory (blue), visual (green), and AV (red) non-sense syllables. (Panel **A**) Scalp distribution of auditory ERPs to auditory, visual and AV speech presentation. (Panel **B**) Latency (bottom left) and absolute amplitude (bottom right) differences of the auditory ERPs (N1 is blue, P2 is red) as a function of correct identification (CI) of visual speech. The better the identification rate in visual speech alone, the earlier the N1/P2 complex occurred. A similar amplitude decrease for N1 (less negative) and P2 (less positive) was observed for all congruent and incongruent AV presentations as compared to A presentations (van Wassenhove et al., 2005).

Wassenhove et al., 2005; Poeppel et al., 2008; Arnal et al., 2009, 2011).

TEMPORAL ENCODING WINDOWS AND TEMPORAL WINDOWS OF INTEGRATION

Two features of the AbyS model are of particular interest here (**Figure 5**). First, visual speech is argued to predict auditory speech in part because of the natural precedence of incoming visual speech inputs; second, AV speech integration tolerates large AV asynchronies without affecting optimal integration (Massaro et al., 1996; Conrey and Pisoni, 2006; van Wassenhove et al., 2007; Maier et al., 2011). In one of these studies (van Wassenhove et al., 2007), two sets of AV speech stimuli (voiced and voiceless auditory bilabials dubbed onto visual velars) were desynchronized and tested using two types of task: (i) a speech identification task ("what do you hear while looking at the talking face?") and (ii) a temporal synchrony judgment task ("where AV stimuli in- or out-of-sync?). Results showed that both AV speech identification and temporal judgment tolerated about 250 ms of AV desynchrony in McGurked and congruent syllables. The duration of the "temporal window of integration" found in these experiments approximated the average syllabic duration across languages, suggesting that syllables may be an important unit of computations in AV speech processing. Additionally, this temporal window of integration showed an asymmetry so that visual leads were better tolerated than auditory leads—with respect to the strength of AV integration. This suggested that the temporal resolutions for the processing of speech information arriving in each sensory modality may actually differ, in agreement with the natural sampling strategies found in auditory and visual systems. This interpretation could now be refined (**Figure 4**).

The "temporal window of integration" can be seen as the integration of two temporal encoding windows (following the precise specifications of Theunissen and Miller, 1995), namely: the encoding window needed by the auditory system to reach phonological categorization is determined by the tolerance to

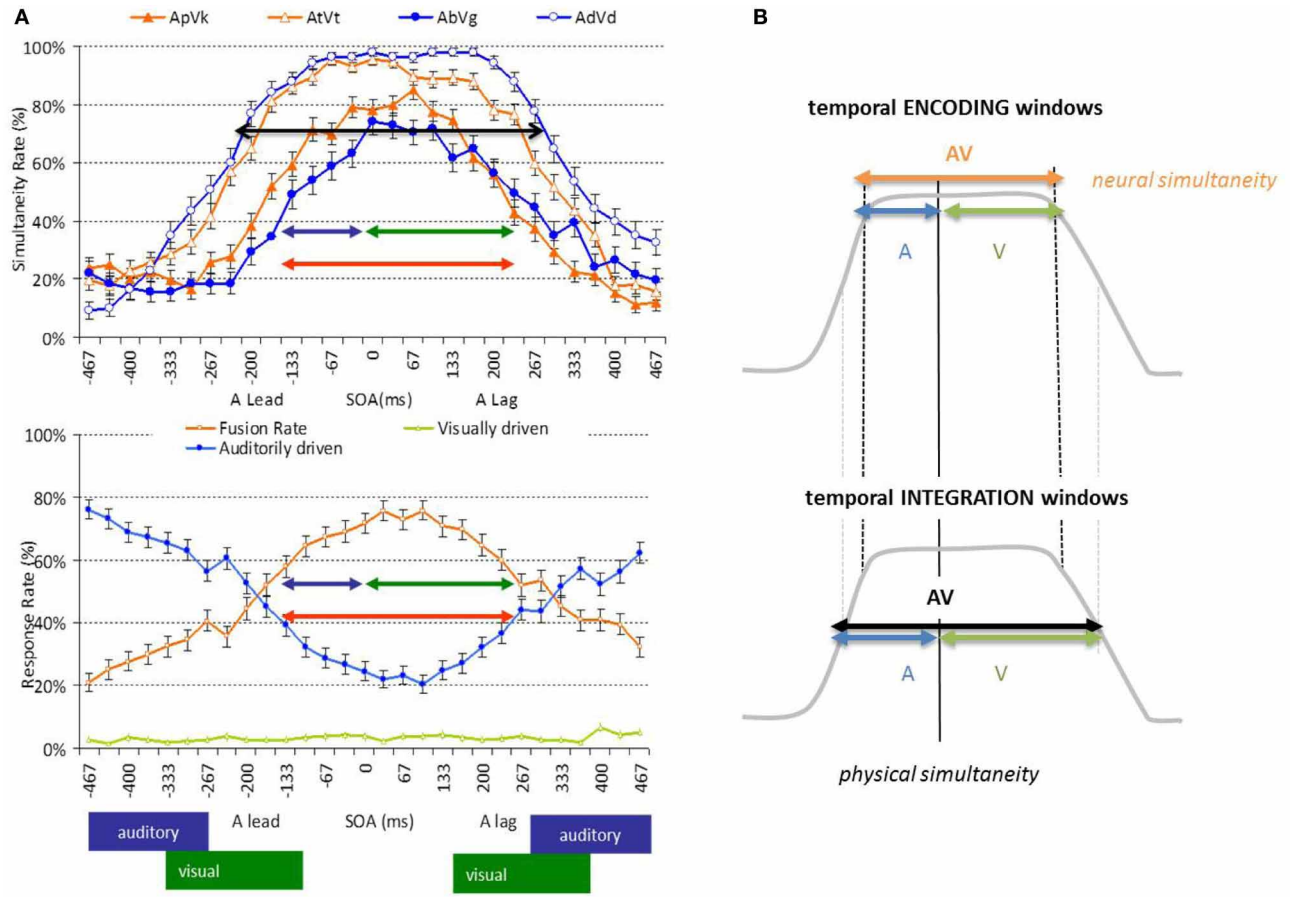


FIGURE 4 | Temporal window of integration in AV speech. (Panel **A**) Illustration of results in a simultaneity judgment task (top) and a speech identification task (bottom) (van Wassenhove et al., 2007). Simultaneity ratings observed for congruent (top, filled symbols) and incongruent (top, open symbols) AV speech as a function of AV desynchrony. Auditory dominated (bottom, blue), visual dominated (bottom, green) or McGurk fusion (bottom, orange) responses as a function of desynchrony using McGurked syllables. The combination of the auditory encoding (blue arrow: tolerance to visual lags) and visual encoding (green arrow: tolerance to visual leads) form the temporal encoding window for AV speech integration. **(**Panel **B)** Schematic illustration distinguishing temporal encoding and temporal integration windows. The temporal resolution reflected in the encoding window corresponds to the necessary or obligatory time for speech encoding; the temporal resolution reflected in the integration windows correspond to the encoding window plus the tolerated temporal noise leading to less than optimal encoding performance.

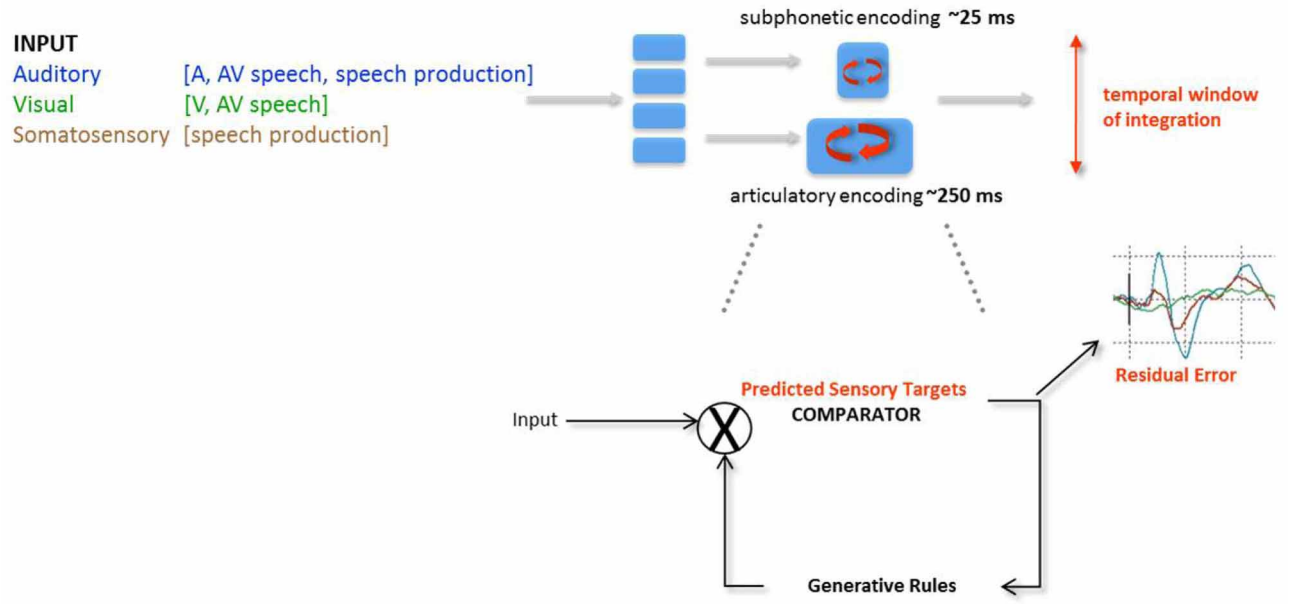


FIGURE 5 | Analysis-by-synthesis (AbyS) in AV speech processing. Two analytical routes are posited on the basis of the original AbyS proposal, namely a subphonetic feature and an articulatory analysis of incoming speech inputs. The privileged route for auditory processing is subphonetic by virtue of the fine temporal precision afforded by the auditory system; the privileged route for visual speech analysis is articulatory by virtue of slower temporal resolution of the visual system and the kinds of information provided by the interlocutor's face. Evidence for the coexistence of 2 modes of speech processing or temporal multiplexing of AV speech can be drawn from the asymmetry of the temporal window of integration in AV speech (cf. **Figure 4**). Although both stages are posited to run in parallel, predictions in both streams are elaborated on the basis of the generative rules of speech production. Predictive mode of AV speech processing is notably marked by a decreased amplitude of the auditory evoked responses (van Wassenhove et al., 2005; Arnal et al., 2009) and residual errors have been characterized either by latency shifts of the auditory evoked responses commensurate with the gain of information in visual speech (van Wassenhove et al., 2005) or by later amplitudes differences commensurate to the detected incongruency of auditory and visual speech inputs (Arnal et al., 2009). AbyS is thus a predictive model operating on temporal multiplexing of speech (i.e., parallel and predictive processing of speech features on two temporal scales) and is compatible with recently proposed neurophysiological implementations of predictive speech coding (Poeppel, 2003; Giraud and Poeppel, 2012).

visual speech lags, whereas the encoding window needed for the visual system to reach visemic categorization is illustrated by the tolerance to auditory speech lags. Hence, the original "temporal window of integration" is a misnomer: the original report describing a plateau within which the order of auditory and speech information did not diminish the rate of integration specifically illustrates the "temporal encoding window" of AV speech i.e., *the necessary time needed for the speech system to elaborate a final outcome or to establish a robust residual error from the two analytical streams in the AbyS framework*. The tolerated asynchronies measured by just-noticeable-differences (Vroomen and Keetels, 2010) or thresholds should be interpreted as the actual "temporal integration window" namely, the tolerance to temporal noise in the integrative system. Said differently, *the fixed constraints are the temporal encoding windows; the tolerance to noise is reflected in the temporal integration windows*.

Temporal windows of integration or "temporal binding windows" (Stevenson et al., 2012) have been observed for various AV stimuli and prompted some promising models for the integration of multisensory information (Colonius and Diederich, 2004). Consistent with the distinction between encoding and integration windows described above, a refined precision of temporal integration/binding windows can be obtained after training (Powers et al., 2009) with a likely limitation of training to the temporal encoding resolution of the system. Interestingly, a recent study (Stevenson et al., 2012) has shown that the width of an individual's temporal integration window for non-speech stimuli could predict the strength of AV speech integration (Stevenson et al., 2012). Whether direct inferences can be drawn between the conscious simultaneity of AV events (overt comparison of events timing entails segregation) and AV speech (integration of AV speech content) is, however, growing controversial. For instance, temporal windows in patients with schizophrenia obtained in a timing task are a poor predictors of their ability to bind AV speech information (Martin et al., 2012), suggesting that distinct neural processes are implicated in the two tasks (in spite of identical AV speech stimuli). Future work in the field will likely help disambiguating which neural operations are sufficient and necessary for conscious timing and which are necessary for binding operations.

OSCILLATIONS AND TEMPORAL WINDOWS

In this context, one could question whether the precedence of visual speech is a prerequisite for predictive coding in AV speech and specifically, whether the ordering of speech inputs in each sensory modality may affect the posited predictive scheme. This

would certainly be an issue if speech analysis followed serial computations operating on a very refined temporal grain. As seen in studies of desynchronized AV speech, this does not seem to be the case: the integrative system operates on temporal windows within which order is not essential (cf. van Wassenhove, 2009 for a discussion on this topic) and both auditory and visual systems likely use different sampling rates in their acquisition of sensory evidence (cf. Temporal parsing and non-invariance).

Recent models of speech processing have formulated clear mechanistic hypotheses implicating neural oscillations: the temporal logistics of cortical activity naturally impose temporal granularities on the parsing and the integration of speech information (Giraud and Poeppel, 2012). For instance, the default oscillatory activity observed in the speech network (Morillon et al., 2010) is consistent with the posited temporal multiplexing of speech inputs. If the oscillatory hypothesis is on the right track, it is thus very unlikely that the dynamic constraints as measured by the temporal encoding (and not integration) window can be changed considering that cortical rhythms (Wang, 2010) provide the dynamic architecture for neural operations. The role of oscillations for predictive operations in cortex has further been reviewed elsewhere (Arnal and Giraud, 2012).

Additionally, visual speech may confer a natural rhythmicity to the syllabic parsing of auditory speech information (Schroeder et al., 2008; Giraud and Poeppel, 2012) and this could be accounted for by phase-resetting mechanisms across sensory modalities. Accordingly, recent MEG work illustrates phase consistencies during the presentation of AV information (Luo et al., 2010; Zion Golumbic et al., 2013). Several relevant oscillatory regimes [namely theta (4 Hz, ∼250 ms), beta (∼20 Hz, 50 ms) and gamma (>40 Hz, 25 ms)] have also been reported that may constrain the integration of AV speech (Arnal et al., 2011). A bulk of recent findings provides structuring constraints on speech processing—i.e., fixed priors. Consistent with neurophysiology, AbyS incorporates temporal multiplexing for speech processing thereby parallel temporal resolutions are used to represent relevant speech information at the segmental and syllabic scales (Poeppel, 2003; Poeppel et al., 2008). In AV speech, each sensory modality may thus operate with a preferred temporal granularity and it is the integration of the two processing streams that effectively reflects the temporal encoding window. Such parallel encoding may also be compatible with recent efforts in modeling AV speech integration (Altieri and Townsend, 2011).

CRITICAL PERIOD IN AV SPEECH PERCEPTION: ACQUISITION OF FIXED PRIORS

During development, the acquisition of speech production could undergo an imitative stage from visual speech perception to speech production. In principle, the imitative stage allows children to learn how to articulate speech sounds by explicitly reproducing the caretakers' facial gestures. However, mounting evidence suggests that imitation does not operate on a blank-slate system; rather, internal motor representations for speech are readily available early on. First, the gestural repertoire is already very rich only 3 weeks after birth, suggesting an innate ability for the articulation of elementary speech sounds (Meltzoff and Moore, 1979; Dehaene-Lambertz and DehaeneHertz-Pannier, 2002). Second, auditory inputs alone are sufficient for infants to reproduce accurately simple speech sounds and enable the recognition of visual speech inputs matching utterances that have only been heard (Kuhl and Meltzoff, 1982, 1984). Furthermore, during speech acquisition, infants do not see their own gestures: consequently, infants can only correct their own speech production via auditory feedback or via matching a peer's gestures (provided visually) to their own production, i.e., via proprioception (Meltzoff, 1999).

Comparatively few studies have addressed the question of AV speech processing during development. The simplest detection of AV synchrony has been argued to emerge first followed by duration, rate and rhythm matching across sensory modalities in the first 10 months of an infant's life (Lewkowicz, 2000). In the spatial domain, multisensory associations are established slowly during the first 2 years of life suggesting that the more complex the pattern, the later the acquisition, in agreement with the "increasing specificity hypothesis" (Gibson, 1969; Spelke, 1981). Three and a half months old infants are sensitive to natural temporal structures but only later on (7 months) are arbitrary multisensory associations detected (e.g., pitch and shape Bahrick, 1992); emotion matching in strangers (Walker-Andrews, 1986). However, early sensitivity to complex AV speech events has been reported in 5 months old infants who can detect the congruency of auditory speech inputs with facial articulatory movements (Rosenblum et al., 1997). The spatiotemporal structuring of arbitrary patterns as well as the nature and ecological relevance of incoming information owe to be important factors in the tuning of a supramodal system. The acquisition of cross-sensory equivalences seems to undergo a perceptual restructuring that can be seen as a fine-tuning of perceptual grouping (Gestalt-like) rules.

Born deaf children who received implants at various ages provide an opportunity to investigate the importance of age at the time of implant for the development of AV speech perception (Bergeson and Pisoni, 2004). A substantial proportion of children who receive cochlear implants learn to perceive speech remarkably well using their implants (Waltzman et al., 1997; Svirsky et al., 2000; Balkany et al., 2002) and are able to integrate congruent AV speech stimuli (Bergeson et al., 2003, 2005; Niparko et al., 2010). In a previous study (Schorr et al., 2005), born-deaf children who had received cochlear implants were tested with McGurk stimuli [visual [ka] dubbed with auditory [pa]; (McGurk and MacDonald, 1976)]. The main hypothesis was that experience played a critical role in forming AV associations for speech perception. In this study, most children with cochlear implants did not experience reliable McGurk effects, and AV speech perception for these children was essentially dominated by lip-reading consistent with their hearing-impairment. However, the likelihood of consistent McGurk illusory reports depended on the age at which children received their cochlear implants. Children who exhibited consistent McGurk illusions received their implants before 30 months of age; conversely, children who received implants after 30 months of age did not show consistent McGurk effects. These results demonstrated that AV speech integration was shaped by experience early on in

life. When auditory experience with speech was mediated by a cochlear implant, the likelihood of acquiring strong AV speech fusion was greatly increased. These results suggested the existence of a sensitive period for AV speech perception (Sharma et al., 2002).

To date however, whether the temporal constraints and neurophysiological indices for AV speech integration in development are comparable to those observed in adults remain unclear.

RESILIENT TEMPORAL INTEGRATION AND THE CO-MODULATION HYPOTHESIS

In natural scenes, diverse sensory cues help the brain select and integrate relevant information to build internal representations. In the context of perceptual invariance and supramodal processing, auditory pitch and visual spatial frequency have been shown to undergo automatic cross-sensory matching (Maeda et al., 2004; Evans and Treisman, 2010). Additionally, auditory and visual signals showing slow temporal fluctuations are most likely to undergo automatic integration (Kösem and van Wassenhove, 2012). In AV speech, the acoustic envelope and the movements of the lips show high correlation or co-modulation (Grant and Seitz, 2000; Remez, 2003) naturally locked to the articulatory gestures of the face. Crucially, this co-modulation shows specificity: AV speech intelligibility shows a similar range of tolerance to asynchronies when the spectral characteristics of the acoustic signal preserve the feature information specific to the articulation (i.e., the F2/F3 formants region) (Grant and Greenberg, 2001). These local correlations have recently been argued to promote AV speech integration even when visual speech information is consciously suppressed (Alsius and Munhall, 2013). Taken altogether, these results suggest that the correlation of auditory and visual speech signals serve as a strong (bottom-up) cue for integration enabling the brain to correctly track signals belonging to the same person as indicated by recent neurophysiological findings (Zion Golumbic et al., 2013).

These observations need to be reconciled with an efficient predictive coding framework as the speech content provided by audition and vision is likely undergoing a non-correlative operation. This would be necessary to allow for the typical informational gain observed in AV speech studies in line with a previously sketched out idea (van Wassenhove et al., 2005), the proposed distinction between correlated and complementary modes of AV speech processing (Campbell, 2008) and AV speech integration models (Altieri and Townsend, 2011).

In this context, while there is ample evidence that speaking rate has a substantial impact on AV speech perception, little is known about the effect of speaking rate on the temporal encoding window. Changes in speaking rate naturally impact the kinematics of speech production, hence the acoustic and visual properties of speech. It is unclear to which extent the posited hard temporal constraints on AV speech integration may be flexible under various speaking rates. In the facial kinematics, different kinds of cues can effectively vary including the motion of the surface structures, the velocity patterns of the articulators and the frequency components over a wide spectrum. Any or all of these could contribute differently to AV speech integration for fast and slow speech and could thus perturb the integration process.

In two experiments (Brungart et al., 2007, 2008), the resilience of AV speech intelligibility was put to the test of noise, AV speech asynchrony and speaking rate. In a first experiment, AV speech recordings of phrases from the Modified Rhyme Test (MRT) were accelerated or decelerated (Brungart et al., 2007). Eight different levels of speaking rate were tested ranging from 0.6 to 20 syllables per second (syl/s). Results showed that the benefits of AV speech were preserved at speaking rates as fast as 12.5 syl/s but disappeared when the rate was increased to 20 syl/s. Importantly, AV speech performance did not benefit from phrases presented slower than their original speaking rates. Using the same experimental material, both the speaking rate and the degree of AV speech asynchrony were varied (Brungart et al., 2008). For the fastest speaking rates, maximal AV benefit occurred at slightly larger visual delay (150 ms) but there was no conclusive evidence suggesting that auditory speech delays for maximal benefit systematically changed with speaking rate. At the highest speaking rates, AV speech enhancement was maximal when the audio signal was delayed by ∼150 ms relative to visual speech, and performance degraded relatively rapidly when the audio speech varied away from its optimal value. As the speaking rate decreased, the range of delays for enhanced AV speech benefit increased, suggesting that participants were tolerant to a wider range of AV speech asynchronies when the speaking rate was relatively slow. However, there was no compelling evidence suggesting that the optimal delay value for AV enhancement systematically changed with the speaking rate of the talker. Finally, when acoustic noise was added, the benefit of visual cues degraded rapidly with faster speaking rate. AV speech integration in noise occurred at all speaking rates slower than 7.8 syl/s. AV speech benefits were observed in all conditions suggesting that the co-modulation of AV speech information can robustly drives integration.

NEURAL MECHANISMS FOR AV SPEECH PROCESSING: CONVERGENCE AND DIVERGENCE

Two reliable electrophysiological markers for AV speech integration are (i) an amplitude decrease (Besle et al., 2004; Jääskeläinen et al., 2004; van Wassenhove et al., 2005; Bernstein et al., 2008; Arnal et al., 2009; Piling, 2009) and (ii) latency shifts (van Wassenhove et al., 2005; Arnal et al., 2009) of the auditory evoked responses. Decreased amplitude of the auditory response to visual speech inputs was originally observed when participants were shown with a video of a face articulating the same or a different vowel sound 500 ms after the presentation of the face (Jääskeläinen et al., 2004). In this study, visual speech inputs were interpreted as leading to the adaptation of the subset of auditory neurons responsive to that feature. However, no difference in amplitude was observed when the visual stimuli were drawn from the same or from a different phonetic category, suggesting non-specific interactions of visual speech information with the early auditory analysis of speech. The amplitude reduction of the auditory evoked responses observed in EEG and MEG is supported by intracranial recordings (Reale et al., 2007; Besle et al., 2008). In particular, Besle et al. (2008) reported two kinds of AV interactions in the secondary auditory association cortices after the first influence of visual speech in this region: at the onset of the

auditory syllable, the initial visual influence disappeared and the amplitude of the auditory response decreased compared to the auditory alone presentation. Similar amplitude reductions were observed to the presentation of AV syllables over the left lateral pSTG (Reale et al., 2007).

In all of these studies, the reported amplitude reduction spanned a couple hundreds of milliseconds, consistent with the implication of low frequency neural oscillations. In monkey neurophysiology, a decreased low-frequency power in auditory cortex has been reported in the context of AV communication (Kayser and Logothetis, 2009). Based on a set of neurophysiological recordings in monkeys, it was proposed that visual inputs change the excitability of auditory cortex by resetting the phase of ongoing oscillation (Schroeder et al., 2008); recent evidence using an AV cocktail party design (Zion Golumbic et al., 2013) support this hypothesis. Additional MEG findings suggest that the tracking of AV speech information may be dealt with by phase-coupling of auditory and visual cortices (Luo et al., 2010). In the context of a recent neurocomputational framework for speech processing (Giraud and Poeppel, 2012), visual speech would thus influence ongoing auditory activity so as to condition the analysis of auditory speech events. Whether this tracking is distinctive with regards to speech content is unclear. The decreased amplitude of auditory evoked responses may be related to the phase entrainment between auditory and visual speech or to the power decrease of low-frequency regions. However, since no clear correlation between the amplitude and the phonetic content are seen in the amplitude, this mechanism does not appear to carry the content of the speech representation, consistent with the lack of visemic or AV speech congruency effect (van Wassenhove et al., 2005; Arnal et al., 2009) and a previously emitted interpretation (Arnal et al., 2009, 2011).

With respect to latency shifts, two studies reported auditory evoked responses as a function of visemic information: one study interpreted that effects on auditory evoked responses carried the residual error (van Wassenhove et al., 2005) and another reported late residual errors at about 400 ms (Arnal et al., 2009). The specificity of this modulation remains unsettled: visual inputs have been reported to change the excitability of auditory cortex by resetting the phase of ongoing oscillation (Lakatos et al., 2008) but an amplification of the signal would have been predicted in auditory cortex (Schroeder et al., 2008). A recent study (Zion Golumbic et al., 2013) implicates the role of attention in selecting or predicting relevant auditory inputs on the basis of visual information. This interpretation would be in line with the notion that visual speech information enables to increase the salience of relevant auditory information for further processing. To which extent phase-resetting mechanisms are speech-specific or more generally implicated in modulating the gain of sensory inputs remains to be determined, along with the implication of specific frequency regimes. Recent findings suggest that multiplexing of speech features could be accomplished in different frequency regimes (Arnal et al., 2011) with coupling between auditory and visual cortices realized via STS. The directionality of these interactions remains to be thoroughly described in order to understand how specific the informational content propagates in the connectivity of these regions. Recent work in monkey neurophysiology has started addressing these issues (Kayser et al., 2010; Panzeri et al., 2010).

It is noteworthy that MEG, EEG, and surface EEG (sEEG) data can contrast with fMRI and PET findings in which enhanced and supra-additive BOLD activations have been reported to the presentation of visual and AV speech. Both enhanced and sub-additive activation in mSTG, pSTG and pSTS have been reported together with left inferior temporal gyrus (BA 44/45), premotor cortex (BA 6), and anterior cingulate gyrus (BA 32) to the presentation of congruent and incongruent AV speech, respectively (Calvert, 1997; Calvert et al., 1999, 2000; Hasson et al., 2007; Skipper et al., 2007). Other fMRI findings (Callan et al., 2003) have shown significant activation of the MTG, STS, and STG in response to the presentation of AV speech in noise; BOLD activation consistent with the inverse effectiveness principle in these same regions (MTG, STS, and STG) has also been reported for stimuli providing information on the place of articulation (Callan et al., 2004). The left posterior STS has been shown sensitivity to incongruent AV speech (Calvert et al., 2000; Wright et al., 2003; Miller and D'Esposito, 2005). Using fMRI and PET, Sekiyama et al. (2003) used the McGurk effect with two levels of auditory noise; comparison between the low and high SNR conditions revealed a left lateralized activation in the posterior STS and BA 22, thalamus, and cerebellum. However, not all studies support the inverse effectiveness principle in auditory cortex (Calvert et al., 1999; Jones and Callan, 2003). Desynchronizing AV McGurk syllables does not significantly affect activation of the STS or auditory cortex (Olson et al., 2002; Jones and Callan, 2003) whereas others report significant and systematic activation of HG as a function of desynchrony (Miller and D'Esposito, 2005). Recent fMRI studies have reported specialized neural populations in the Superior Temporal Sulcus (STS in monkey) or Superior Temporal Cortex (STC, human homolog). The organization of this multisensory region is known to be patchy (Beauchamp et al., 2004) but recognized to be an essential part of the AV speech integration network (Arnal et al., 2009; Beauchamp et al., 2010). The middle STC (mSTC) is a prime area for the detection of AV asynchrony and the integration of AV speech (Bushara et al., 2001; Miller and D'Esposito, 2005; Stevenson et al., 2010, 2011). At least two neural subpopulations may coexist in this region: the synchrony population tagged S-mSTC showing increased activation to AV speech stimuli when the auditory and visual streams are in synchrony and the bimodal population tagged B-mSTC showing the opposite pattern, namely a decrease of activation with the presentation of synchronized audiovisual speech streams (Stevenson et al., 2010, 2011). These results may help shed light on the contribution of neural subpopulations in mSTC in computing redundant information vs. efficient coding for AV speech processing.

Using fMRI technique, the contribution of motor cortices has also been tested in the perception of auditory, visual and AV speech (Skipper et al., 2007). In these experiments, participants actively produced syllables or passively perceived auditory, visual and AV stimuli in the scanner. The AV stimuli consisted of both congruent AV [pa], [ka], and [ta] and McGurk fusion stimuli (audio [pa] dubbed onto a face articulating [ka]). The main results showed that the cortical activation pattern during

the perception of visual and AV but not auditory speech greatly overlapped with that observed in speech production. The areas showing above 50% of overlap in production and perception were bilateral anterior and posterior Superior Temporal cortices (STa and STp, respectively), and ventral premotor cortex (PMv). The perception of McGurk fusion elicited patterns of activation that correlated differently across cortical areas with the perception of a congruent AV [pa] (the auditory component in the McGurk fusion stimulus), AV [ka] (the visual component of the McGurk fusion stimulus) or AV [ta] (the perceived illusory [ta] elicited by the McGurk fusion stimulus). Activations observed in frontal motor areas, and auditory and somatosensory cortices during McGurk presentation correlated more with the perceived syllable (AV [ta]) than the presented syllables in either sensory modality (A [pa], V [ka]). In visual cortices, activation correlated most with the presentation of a congruent AV [ka]. Overall, results were interpreted in the context of a forward model of speech processing.

OUTSTANDING QUESTIONS

First, what is (in) a prediction? Although computational models provide interesting constraints with which to work, we cannot currently separate temporal prediction from speech-content predictions (e.g., Arnal and Giraud, 2012). One important finding encompassing the context of speech is that amplitude decrease can be seen as a general marker of predictive coding (e.g., Todorovic and de Lange, 2012) in auditory cortex and more specifically during speech production (Houde and Jordan, 1998).

Second, what anchors are used to parse visual speech information or, what are the "edges" (Giraud and Poeppel, 2012) of visual speech information? Complementarily, can we use cortical responses to derive the distinctive features of visual speech (Luo et al., 2010)?

Third, in the context of fixed temporal constraints for speech processing, how early can temporal encoding/integration windows be characterized in babies? Is the co-modulation hypothesis a general guiding principle for multisensory integration or a specific feature of AV speech?

Finally, the implication of the motor system in the analysis of speech inputs has been a long-standing debate undergoing increasing refinement (e.g., Scott et al., 2009). The inherent rhythmicity of speech production naturally constrains the acoustic and visual structure of auditory and visual speech outcomes. Is the primary encoding mode of AV speech articulatory or acoustic (e.g., Altieri et al., 2011; Schwartz et al., 2012)?

ACKNOWLEDGMENTS

This work was realized thanks to a Marie Curie IRG-24299, an ERC-StG-263584 and an ANR10JCJ-1904 to Virginie van Wassenhove.

REFERENCES

Allik, J., and Konstabel, K. (2005). G. F. Parrot and the theory of unconscious inferences. *J. Hist. Behav. Sci.* 41, 317–330. doi: 10.1002/jhbs.20114

Alsius, A., and Munhall, K. G. (2013). Detection of audiovisual speech correspondences without visual awareness. *Psychol. Sci.* 24, 423–431. doi: 10.1177/0956797612457378

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

Altieri, N., Pisoni, D. B., and Townsend, J. T. (2011). Some behavioral and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. *Seeing Perceiving* 24, 513–539. doi: 10.1163/187847611X595864

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

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., Morillon, B., Kell, C., and Giraud, A. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

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

Auer, E. J. (2002). The influence of the lexicon on speech read word recognition: contrasting segmental and lexical distinctiveness. *Psychon. Bull. Rev.* 9, 341–347. doi: 10.3758/BF03196291

Bahrick, L. E. (1992). Infant's perceptual differentiation of amodal and modaliy-specific audio-visual realations. *J. Exp. Child Psychol.* 53, 180–199. doi: 10.1016/0022-0965(92)90048-B

Balkany, T. J., Hodges, A. V., Eshraghi, A. A., Butts, S., Bricker, K., Lingvai, J., et al. (2002). Cochlear implants in children-a review. *Acta Otolaryngol.* 122, 356–362. doi: 10.1080/00016480260000012

Barlow, H. (1961). "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, ed W. Rosenblith (Cambridge: MIT Press), 217–234.

Barlow, H. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Res.* 30, 1561–1571. doi: 10.1016/0042-6989(90)90144-A

Barlow, H., and Földiak, P. (1989). "Adaptation and decorrelation in the cortex," in *The Computing Neuron*, eds R. Durbin, C. Miall, and G. Mitchison (Wokingham: Addison-Wesley), 54–72.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., and Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J. Cogn. Neurosci.* 17, 377–391. doi: 10.1162/0898929053279586

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., and Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.* 7, 1190–1192. doi: 10.1038/nn1333

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

Bergeson, T. R., and Pisoni, D. B. (2004). "Audiovisual speech perception in deaf adults and children following cochlear implantation," in *Handbook of Multisensory Integration,* eds G. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: MIT Press), 749–772.

Bergeson, T. R., Pisoni, D. B., and Davis, R. A. (2003). A longitudinal study of audiovisual speech perception by children with hearing loss who have cochlear implants. *Volta Rev.* 103, 347–370.

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

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

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., and Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *J. Neurosci.* 28, 14301–14310. doi: 10.1523/JNEUROSCI.2875-08.2008

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

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 445–463. doi: 10.1037/0096-1523.30.3.445

Brungart, D., Iyer, N., Simpson, B., and van Wassenhove, V. (2008). "The effects of temporal asynchrony on the intelligibility of accelerated speech," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, (Moreton Island, QLD: Tangalooma Wild Dolphin Resort).

Brungart, D., van Wassenhove, V., Brandewie, E., and Romigh, G. (2007). "The effects of temporal acceleration and deceleration on auditory-visual speech perception," in *International Conference on Auditory-Visual Speech Processing (AVSP)* (Hilvarenbeek).

Busch, N. A., and VanRullen, R. (2010). Spontaneous EEG oscillations reveal periodic sampling of visual attention. *Proc. Natl. Acad. Sci. U.S.A.* 107, 16048–16053. doi: 10.1073/pnas.1004801107

Bushara, K. O., Grafman, J., and Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *J. Neurosci.* 21, 300–304.

Callan, D. E., Jones, J. A., Munhall, K. G., Kroos, C., Callan, A. M., and Vaitikiosis-Bateson, E. (2004). Multisensory integrtaion sites identified by perception of spatial wavelet filtered visual speech gesture information. *J. Cogn. Neurosci.* 16, 805–816. doi: 10.1162/089892904970771

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

Calvert, G. A. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 893–596. doi: 10.1126/science.276.5312.593

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

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

Calvert, G. A., and Thesen, T. (2004). Multisensory integratio: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

Campbell, R. (1986). Face recognition and lipreading. *Brain* 109, 509–521. doi: 10.1093/brain/109.3.509

Campbell, R. (1989). "Lipreading," in *Handbook of Research on Face Processing,* eds A. W. Young and H. D. Ellis (Malden: Blackwell Publishing), 187–233.

Campbell, R. (1992). "Lip-reading and the modularity of cognitive function: neuropsychological glimpses of fractionation from speech and faces," in *Analytic Approaches to Human Cognition,* eds J. Alegria, D. Holender, J. Junca de Morais, and M. Radeau (Amsterdam: Elsevier Science Publishers), 275–289.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1001–1010. doi: 10.1098/rstb.2007.2155

Campbell, R., Garwood, J., Franklinavi, 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. doi: 10.1016/0028-3932(90)90003-7

Campbell, C., and Massaro, D. W. (1997). Perception of visible speech: influence of spatial quantization. *Perception* 26, 627–644. doi: 10.1068/p260627

Cathiard, M.-A., and Abry, C. (2007). "Speech structure decisions from speech motion coordinations," in *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saarbrücken.

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

Chomsky, N. (2000). "Recent contributions to the theory of innate ideas," in *Minds, Brains and Computers The foundation of Cognitive Science, an Anthology*, eds R. M. Harnish and D. D. Cummins (Malden, MA: Blackwell), 452–457.

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English*. New York; Evanston; London: Harper and Row.

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

Colonius, H., and Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J. Cogn. Neurosci.* 16, 1000–1009. doi: 10.1162/0898929041502733

Conrey, B., and Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and non speech signals. *J. Acoust. Soc. Am.* 119, 4065. doi: 10.1121/1.2195091

Czigler, I., Winkler, I., Pató, L., Várnagy, A., Weisz, J., and Balázs, L. (2006). Visual temporal window of integration as revealed by the visual mismatch negativity event-related potential to stimulus omissions. *Brain Res.* 1104, 129–140. doi: 10.1016/j.brainres.2006.05.034

de Gelder, B., Böcker, K. B. E., Tuomainen, J., Hensen, M., and Vroomen, J. (1999). The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. *Neurosci. Lett.* 260, 133–136. doi: 10.1016/S0304-3940(98)00963-X

Dehaene-Lambertz, G., S. Dehaene, and Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science* 298, 2013–2015. doi: 10.1126/science.1077066

Denève, S., and Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *J. Neurophysiol. Paris* 98, 249–258. doi: 10.1016/j.jphysparis.2004.03.011

Desimone, R., and Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Res.* 178, 363–380. doi: 10.1016/0006-8993(79)90699-1

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

Erber, M. P. (1978). Auditory-visual speech perception of speech with reduced optical clarity. *J. Speech Hear. Res.* 22, 213–223.

Ernst, M. O., and Bülthoff, H. H. (2004). Meging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169. doi: 10.1016/j.tics.2004.02.002

Evans, K. K., and Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *J. Vis.* 10:6. doi: 10.1167/10.1.6

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

Ghazanfar, A. A., Chandrasekaran, C., and Logothetis, N. K. (2008). Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *J. Neurosci.* 28, 4457–4469. doi: 10.1523/JNEUROSCI.0541-08.2008

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

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

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., and Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* 25:5004. doi: 10.1523/JNEUROSCI.0799-05.2005

Gibson, E. J. (1969). *Principles of Perceptual Learning and Development.* New York, NY: Appleton - Century - Crofts.

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

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: a theoretical perspective. *J. Acoust. Soc. Am.* 112, 30–33. doi: 10.1121/1.1482076

Grant, K. W., and Greenberg, S. (2001). "Speech intelligibility derived from asynchronous processing of auditory-visual information," in *Auditory-Visual Speech Pocessing*, (Aalborg).

Grant, K. W., and Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *J. Acoust. Soc. Am.* 104, 2438–2450. doi: 10.1121/1.423751

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, 1197–1207. doi: 10.1121/1.1288668

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

Green, K. P. (1996). "The use of auditory and visual information in phonetic perception," in *Speechreading by Humans and Machines,* eds D. G. Stork and M. E. Hennecke (Berlin: Springer-Verlag), 55–77.

Greenberg, S. (1998). A syllabic-centric framework for the evolution of spoken language. *Brain Behav. Sci.* 21, 267–268. doi: 10.1017/S0140525X98311176

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

Halle, M., and Stevens, K. N. (1962). Speech recognition: a model and a program for research. *IRE Trans. Inf. Theor.* 8, 155–159. doi: 10.1109/TIT.1962.1057686

Hans-Otto, K. (2001). New insights into the functions of the superior temporal cortex. *Nat. Neurosci.* 2, 568. doi: 10.1038/35086057

Harth, E., Unnnikrishnan, K. P., and Pandya, A. S. (1987). The inversion of sensory processing by feedback pathways: a model of visual cognitive functions. *Science* 237, 184–187. doi: 10.1126/science.3603015

Hasson, U., Skipper, J., Nusbaum, H., and Small, S. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56, 1116–1126. doi: 10.1016/j.neuron.2007.09.037

Hosoya, T., S. A. Baccus, and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature* 436, 71–77. doi: 10.1038/nature03689

Houde, J. F., and Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science* 279, 1213–1216. doi: 10.1126/science.279.5354.1213

Jääskeläinen, I. P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R., et al. (2004). Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *Neuroreport* 18, 2741–2744.

Jones, J., and Callan, D. (2003). Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Neuroreport* 14, 1129–1133. doi: 10.1097/00001756-200306110-00006

Jordan, T. R., McCotter, M. V., and Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. *Percept. Psychophys.* 62, 1394–1404. doi: 10.3758/BF03212141

Kayser, C., and Logothetis, N. K. (2009). Directed interactions between auditory and superior temporal cortices and their role in sensory integration. *Front. Integr. Neurosci.* 3:7. doi: 10.3389/neuro.07.007.2009

Kayser, C., Logothetis, N. K., and Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Curr. Biol.* 20, 19–24. doi: 10.1016/j.cub.2009.10.068

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *J. Neurosci.* 27, 1824. doi: 10.1523/JNEUROSCI.4737-06.2007

Kent, R. D. (1983). "The segmental organization of speech, Chapter 4," in *The Production of Speech*, ed. P. F. MacNeilage (Newyork, NY: Springer-verlag), 57–89.

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4:e1000209. doi: 10.1371/journal.pcbi.1000209

Kihlstrom, J. F. (1987). The cognitive unconscious. *Science* 237, 1445–1452. doi: 10.1126/science.3629249

Kösem, A., and van Wassenhove, V. (2012). Temporal structure in audiovisual sensory selection. *PLoS ONE* 7:e40936. doi: 10.1371/journal.pone.0040936

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

Kuhl, P., and Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behav. Dev.* 7, 361–381. doi: 10.1016/S0163-6383(84)80050-8

Lakatos, P., Karmos, G., Mehta, A., Ulbert, I., and Schroeder, C. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320, 110–113. doi: 10.1126/science.1154735

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* 9, 920–933. doi: 10.1038/nrn2532

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nat. Neurosci.* 5, 356–363. doi: 10.1038/nn831

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychol. Bull.* 162, 281–308. doi: 10.1037/0033-2909.126.2.281

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

Liégeois, C., de Graaf, J. B., Laguitton, V., and Chauvel, P. (1999). Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cereb. Cortex* 9, 484–496. doi: 10.1093/cercor/9.5.484

Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biol.* 8:e1000445. doi: 10.1371/journal.pbio.1000445

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9, 1432–1438. doi: 10.1038/nn1790

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

MacDonald, J., Soren, A., and Bachmann, T. (2000). Hearing by eye: how much spatial degradation can be tolerated. *Perception* 29, 1155–1168. doi: 10.1068/p3020

MacKay, D. M. (1958). Perceptual stability of a stroboscopically lit visual field containing self-luminous objects. *Nature* 181, 507–508. doi: 10.1038/181507a0

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

Maeda, F., Kanai, R., and Shimojo, S. (2004). Changing pitch induced visual motion illusion. *Curr. Biol.* 14, R990–R991. doi: 10.1016/j.cub.2004.11.018

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

Maiste, A. C., Wiens, A. S., Hunt, M. J., Sherg, M., and Picton, T. W. (1995). Event-related potentials and the categorical perception of speech sounds. *Ear Hear.* 16, 68–90. doi: 10.1097/00003446-199502000-00006

Martin, B., Giersch, A., Huron, C., and van Wassenhove, V. (2012). Temporal event structure and timing in schizophrenia: preserved binding in a longer "now". *Neuropsychologia* 51, 358–371. doi: 10.1016/j.neuropsychologia.2012.07.002

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Massaro, D. W. (1998). *Perceiving Talking Faces*. Cambridge: MIT Press.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am.* 100, 1777. doi: 10.1121/1.417342

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

Meltzoff, A. N. (1999). Origins of theory of mind, cognition and communication. *J. Commun. Disord.* 32, 251–226. doi: 10.1016/S0021-9924(99)00009-X

Meltzoff, A. N., and Moore, M. K. (1979). Interpreting "imitative" responses in early infancy. *Science* 205, 217–219. doi: 10.1126/science.451596

Miller, L., 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

Morillon, B., Lehongre, K., Frackowiak, R. S. J., 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., Krause, C., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Brain Res. 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

Murray, M. M., and Spierer, L. (2011). Multisensory integration:

what you see is where you hear. *Curr. Biol.* 21, R229–R231. doi: 10.1016/j.cub.2011.01.064

Näätänen, R. (1995). The mismatch negativity: a powerful tool for cognitive neuroscience. *Ear Hear.* 16, 6–18. doi: 10.1097/00003446-199502000-00002

Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol.* 42, 313–329. doi: 10.1016/0001-6918(78)90006-9

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

Olson, I., Gatenby, J., and Gore, J. (2002). A comparison of bound and unbound audio-visual information processing in the human cerebral cortex. *Brain Res. Cogn. Brain Res.* 14, 129–138. doi: 10.1016/S0926-6410(02)00067-8

Panzeri, S., Brunel, N., Logothetis, N. K., and Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends Neurosci.* 33, 111–120. doi: 10.1016/j.tins.2009.12.001

Paré, M., Richler, R. C., and Ten Hove, M. (2003). Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Percept. Psychophys.* 65, 553–567. doi: 10.3758/BF03194582

Pascual-Leone, A., and Hamilton, R. (2001). The metamodal organization of the brain. *Prog. Brain Res.* 134, 427–445. doi: 10.1016/S0079-6123(01)34028-1

Philips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *J. Cogn. Neurosci.* 12, 1038–1055. doi: 10.1162/08989290051137567

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

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., 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

Powers, A. R., 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

Puce, A., Allison, T., Bentin, A., Gore, J. C., and McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *J. Neurosci.* 18, 2188–2199.

Pylyshyn, Z. (1984). *Computation and Cognition: Towards a Foundation for Cognitive Science*. Cambridge: MIT Press.

Rajkai, C., Lakatos, P., Chen, C., Pincze, Z., Karmos, G., and Schroeder, C. (2008). Transient cortical excitation at the onset of visual fixation. *Cereb. Cortex* 18, 200–209. doi: 10.1093/cercor/bhm046

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Reale, R., Calvert, G., Thesen, T., Jenison, R., Kawasaki, H., Oya, H., et al. (2007). Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience* 145, 162–184. doi: 10.1016/j.neuroscience.2006.11.036

Remez, R. (2003). Establishing and maintaining perceptual coherence: unimodal and multimodal evidence. *J. Phon.* 31, 293–304. doi: 10.1016/S0095-4470(03)00042-1

Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D., and Rubin, P. E. (1998). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Commun.* 26, 65–73. doi: 10.1016/S0167-6393(98)00050-8

Rosen, S. (1992). temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. Lond. B* 336, 367–373. doi: 10.1098/rstb.1992.0070

Rosenblum, L., Schmuckler, M. A., and Johnson, J. A. (1997). The McGurk effect in infants. *Percept. Psychophys.* 59, 347–357. doi: 10.3758/BF03211902

Rosenblum, L. D., and Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percep. Perform.* 22, 318–331. doi: 10.1037/0096-1523.22.2.318

Rosenblum, L., and Yakel, D. A. (2001). The McGurk effect from single and mixed speaker stimuli. *Acoust. Res. Lett. Online* 2, 67–72. doi: 10.1121/1.1366356

Saltzman, E. L., and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* 1, 333–382. doi: 10.1207/s15326969eco0104_2

Sams, M., and Aulanko, R. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–147. doi: 10.1016/0304-3940(91)90914-F

Schorr, E., Fox, N., van Wassenhove, V., and Knudsen, E. (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., 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

Schroeder, C., 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

Schwartz, J., 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 Psycholoy of Speechreading and Auditory-Visual Speech*, eds R. Campbell, B. Dodd, and D. Burnham (East Sussex: Psychology Press), 85–108.

Schwartz, J.-L. (2003). "Why the FLMP should not be applied to McGurk data...or how to better compare models in the Bayesian framework," in *AVSP - International Conference on Audio-Visual Speech Processing*, (St-Jorioz).

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., 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. (1994). Differences in auditory-visual speech perception between Japanese and americans: McGurk effect as a function of incompatibility. *J. Acoust. Soc. Am.* 15, 143–158. doi: 10.1250/ast.15.143

Sekiyama, K. (1997). Cultural and linguistic factors in audio-visual speech processing: the McGurk effect in Chinese subjects. *Percept. Psychophys.* 59, 73–80. doi: 10.3758/BF03206849

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

Servos, P., Osu, R., Santi, A., and Kawato, M. (2002). The neural substrates of biological motion perception: an fMRI study. *Cereb. Cortex* 12, 772–782. doi: 10.1093/cercor/12.7.772

Sharma, A., and Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *J. Acoust. Soc. Am.* 16, 1078–1083. doi: 10.1121/1.428048

Sharma, A., Dorman, M. F., and Spahr, A. J. (2002). Rapid development of cortical auditory evoked potentials after early cochlear implantation. *Neuroreport* 13, 1365–1368. doi: 10.1097/00001756-200207190-00030

Sharma, J., Dragoi, V., and Tenebaum, J. B. (2003). V1 neurons signal acquisition of an internal representation of stimulus location. *Science* 300, 1758–1763. doi: 10.1126/science.1081721

Simos, P. G., Diehl, R. L., Breier, J. I., Molis, M. R., Zouridakis, G., and Papanicolaou, A. C. (1998). MEG correlates of categorical perception of a voice onset time continuum in humans. *Cogn. Brain Res.* 7, 215–219. doi: 10.1016/S0926-6410(98)00037-8

Skipper, J., van Wassenhove, V., Nusbaum, H., and Small, S. (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

Smith, E. C., and Lewicki, M. S. (2006). Efficient auditory coding. *Nature* 439, 978–982. doi: 10.1038/nature04485

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

Spelke, E. S. (1981). The infant's acquisition of knowledge of bimodally specified events. *J. Exp. Child Psychol.* 31, 279–299. doi: 10.1016/0022-0965(81)90018-7

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B Biol. Sci.* 216, 427–459. doi: 10.1098/rspb.1982.0085

Stekelenburg, 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 evidence for a multisensory speech-specific mode of perception. *Neuropsychologia* 50, 1425–1431. doi: 10.1016/j.neuropsychologia.2012.02.027

Stevens, K. (1960). Toward a model of speech perception. *J. Acoust. Soc. Am.* 32, 45–55. doi: 10.1121/1.1907874

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., Van DerKlok, 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., 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

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

Summerfield, A. Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye*, eds B. Dodd and R. Campbell (London: Erlbaum Associates), 3–51.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., and Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychol. Sci.* 11, 153–158. doi: 10.1111/1467-9280.00231

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

Tervaniemi, M., Maury, S., and Näätänen, R. (1994). Neural representations of abstract stimulus features in the human brain as reflected by the mismatch negativity. *Neuroreport* 5, 844–846. doi: 10.1097/00001756-199403000-00027

Theunissen, F., and Miller, J. P. (1995). Temporal encoding in nervous systems: a rigorous definition. *J. Comput. Neurosci.* 2, 149–162. doi: 10.1007/BF00961885

Tiippana, K., Andersen, T. S., and Sams, M. (2003). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

Todorovic, A., and de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *J. Neurosci.* 32, 13389–13395. doi: 10.1523/JNEUROSCI.2227-12.2012

Tuller, B., and Kelso, J. A. (1984). The timing of articulatory gestures: evidence for relational invariants. *J. Acoust. Soc. Am.* 76, 1030–1036. doi: 10.1121/1.391421

Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (2005). Audio-visual speech perception is special. *Cognition* 96, B13–B22. doi: 10.1016/j.cognition.2004.10.004

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.* 98, 11656–11661. doi: 10.1073/pnas.191374198

van Wassenhove, V. (2009). Minding time in an amodal representational space. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 1815–1830. doi: 10.1098/rstb.2009.0023

van Wassenhove, V., Ghazanfar, A., Munhall, K., and Schroeder, C. (2012). "Bridging the gap between human and non human studies of audiovisual integration," in *The New Handbook of Multisensory Processing,* ed B. E. Stein (Cambridge: MIT Press), 153–167.

van Wassenhove, V., Grant, K., 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

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

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., and Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Percept. Psychophys.* 60, 926–940. doi: 10.3758/BF03211929

Viviani, P., Figliozzi, F., and Lacquaniti, F. (2011). The perception of visible speech: estimation of speech rate and detection of time reversals. *Exp. Brain Res.* 215, 141–161. doi: 10.1007/s00221-011-2883-9

Voss, P., and Zatorre, R. J. (2012). Organization and reorganization of sensory-deprived cortex. *Curr. Biol.* 22, R168–R173. doi: 10.1016/j.cub.2012.01.030

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

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20754–20759. doi: 10.1073/pnas.1117807108

Walker, S., Bruce, V., and O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Attent. Percept. Psychophys.* 57, 1124–1133. doi: 10.3758/BF03208369

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: relation of eye and voice. *Dev. Psychol.* 22, 373–377. doi: 10.1037/0012-1649.22.3.373

Waltzman, S. B., Cohen, N. L., Gomolin, L. H., Green, J. E., Shapiro, W. H., Hoffman, R. A., et al. (1997). Open-set speech perception in congenitally deaf children using cochlear implants. *Am. J. Otol.* 18, 342–349.

Wang, X. J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiol. Rev.* 90, 1195–1268. doi: 10.1152/physrev.00035.2008

Werner-Reiss, U., Kelly, K., Trause, A., Underhill, A., and Groh, J. (2003). Eye position affects activity in primary auditory cortex of primates. *Curr. Biol.* 13, 554–562. doi: 10.1016/S0960-9822(03)00168-4

Wright, T., Pelphrey, K., Allison, T., McKeown, M., 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

Wundt, W. (1874). *Grundzuge Derphysiologischen Psychologie*, Leipzig: Engelmann.

Yabe, H., Tervaniemi, M., Reinikainen, K., and Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *Neuroreport* 8, 1971–1974. doi: 10.1097/00001756-199705260-00035

Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis. *Trends Cogn. Sci.* 10, 301–308. doi: 10.1016/j.tics.2006.05.002

Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J. Neurosci.* 33, 1417–1426. doi: 10.1523/JNEUROSCI.3675-12.2013

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: 09 April 2013; paper pending published: 28 April 2013; accepted: 10 June 2013; published online: 12 July 2013.

*Citation: 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*

This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 10 September 2013
doi: 10.3389/fpsyg.2013.00615

Neural dynamics of audiovisual speech integration under variable listening conditions: an individual participant analysis

Nicholas Altieri*[1] *and* ***Michael J. Wenger*[2]**

[1] *Department of Communication Sciences and Disorders, Idaho State University, Pocatello, ID, USA*
[2] *Department of Psychology, The University of Oklahoma, Norman, OK, USA*

Edited by:
Nuria Sebastian-Galles, Universitat Pompeu Fabra, Spain

Reviewed by:
Pia Knoeferle, Bielefeld University, Germany
Agnes Alsius, Queen's University, Canada

***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

Speech perception engages both auditory and visual modalities. Limitations of traditional accuracy-only approaches in the investigation of audiovisual speech perception have motivated the use of new methodologies. In an audiovisual speech identification task, we utilized *capacity* (Townsend and Nozawa, 1995), a dynamic measure of efficiency, to quantify audiovisual integration. Capacity was used to compare RT distributions from audiovisual trials to RT distributions from auditory-only and visual-only trials across three listening conditions: clear auditory signal, S/N ratio of −12 dB, and S/N ratio of −18 dB. The purpose was to obtain EEG recordings in conjunction with capacity to investigate how a late ERP co-varies with integration efficiency. Results showed efficient audiovisual integration for low auditory S/N ratios, but inefficient audiovisual integration when the auditory signal was clear. The ERP analyses showed evidence for greater audiovisual amplitude compared to the unisensory signals for lower auditory S/N ratios (higher capacity/efficiency) compared to the high S/N ratio (low capacity/inefficient integration). The data are consistent with an interactive framework of integration, where auditory recognition is influenced by speech-reading as a function of signal clarity.

Keywords: capacity, integration, multisensory speech, models of integration, Late ERPs, audiovisual integration, audiovisual interactions

Studies of audiovisual speech recognition have revealed the dramatic effect that visual information can have on the processing of auditory speech inputs. One of the most significant findings is that visual speech signals provided by a talker's face enhance identification accuracy, especially when listening conditions become degraded (e.g., Sumby and Pollack, 1954; see Ross et al., 2007). Accuracy data from audiovisual speech identification experiments have pointed to a specific range of auditory signal-to-noise (S/N) ratios in which audiovisual integration occurs most efficiently (Ross et al., 2007). For example, Grant et al. (1998) fit models of consonant identification that allow the relative contribution of each information source to be estimated from the data (see Braida, 1991; Massaro, 2004). The authors applied these models to data sets obtained from normal-hearing and hearing-impaired subjects in identification experiments. These studies indicate considerable individual variability in the ability to combine auditory and visual information. This variability has been observed in both normal-hearing and hearing impaired listeners (see Grant et al., 1998).

The implication of these studies is that the visual signal affords variable levels of integration efficiency under different listening conditions. Specifically, this suggests that integration occurs in fundamentally distinct ways under different auditory S/N ratios and across different populations such as normal-hearing vs. hearing-impaired (e.g., Sommers et al., 2005). Also, an important aspect of speech recognition for both unisensory and multisensory cases concerns the temporal nature of the speech signal. Speech recognition unfolds in real-time, and audiovisual speech studies that do not employ measures of the dynamics of processing can miss important characteristics of neural and cognitive mechanisms (Altieri et al., 2011). A unified approach for investigating audiovisual speech integration must combine real-time behavioral measures with dynamic brain signals (Besle et al., 2004; van Wassenhove et al., 2005, 2007; Pilling, 2009; Cappe et al., 2010). This will involve combining EEG amplitude with model based reaction time (RT) methods (see e.g., Altieri, 2010; Altieri and Townsend, 2011; see also Colonius and Diederich, 2010).

Our study utilizes a combined EEG and RT model-based approach to investigate the following questions: (1) under which listening conditions does visual speech information yield the most efficient integration? (2) At which points in time during speech recognition does the visual signal have the greatest influence on the auditory speech percept? And (3), to what extent are neural measures of efficiency predictive of model based behavioral measures of efficiency? This latter point is especially important because EEG amplitude can indicate neural firing associated with sensory processing, extraction of features, and recognition/categorization. For example, one study using a spoken word recognition test in children with hearing loss observed ERPs of approximately normal amplitude and latency in children with better speech recognition, but significantly reduced or absent

ERPs in those poor word recognition ability (Rance et al., 2002). Nonetheless, ERP studies have almost universally failed to relate ERPs to a quantitative behavioral index of processing ability that makes predictions relative to a well-defined behavioral model (although cf. Winneke and Phillips, 2011).

To address this latter issue, we obtained a behavioral measure of integration efficiency known as *capacity* that uses non-parametric predictions of parallel independent processing models (Townsend and Nozawa, 1995; Altieri and Townsend, 2011) as a benchmark for efficient integration. While we measure capacity/efficiency behaviorally, "capacity" does not directly refer to processing architecture (e.g., parallel vs. coactive; Miller, 1982; Townsend and Nozawa, 1995). Second, we obtained brain recordings to examine the extent to which ERPs systematically covary with capacity across listening conditions.

NEURO-COGNITIVE BASIS OF INTEGRATION EFFICIENCY

Evidence obtained from EEG (e.g., Ponton et al., 2009; Naue et al., 2011; Winneke and Phillips, 2011) as well as RT studies (e.g., Altieri and Townsend, 2011; Altieri et al., 2011) is consistent with the hypothesis that unisensory processing occurs in separate channels, with cross-modal interactions occurring between them (cf. Rosenblum, 2005). In the ERP literature, Winneke and Phillips (2011) used a combination of RTs and ERPs to assess integration skills across age group. The analysis of the RT data and pre-linguistic ERP components associated with auditory-visual detection, revealed early dependencies between sensory modalities. The analysis of the N1/P1 components showed an amplitude reduction in both components on audiovisual trials relative the auditory-only plus visual-only trials. However, the precise physiological relationships between patterns of brain signals and variations in integration efficiency, and the manner in which those co-variations relate to the predictions for cross-modal dependencies have yet to be established.

Moreover, Altieri and Townsend (2011) fit processing models to RT distributions obtained from audiovisual identification data and found that a parallel model, with separate auditory and visual channels and a first terminating (OR) decision rule (see Townsend and Nozawa, 1995; Townsend and Wenger, 2004) best accounted for the data. **Figure 1** shows a schema of this model. First, auditory, and visual speech cues are input, which undergo early sensory processing (prior to conscious language recognition). Subsequently, language based features such as phonemes and visual cues about place of articulation (Summerfield, 1987) are extracted, and information related to a percept is accumulated until a decision bound is reached using an OR decision rule. That is, recognition occurs as soon as either enough auditory or enough visual speech information is available (Altieri and Townsend, 2011). To use an example, suppose that as soon as enough auditory evidence for a word/category, say [Date], reaches threshold, the listener perceives "Date." A critical feature of this model is that we hypothesize that the channels are not independent—hence the arrows showing cross-modal connections. We primarily concern ourselves with audiovisual interactions occurring during linguistic analysis, although evidence exists for earlier interactions. The capacity results will be critical in falsifying null-hypothesis assumptions of independence, pointing instead to dependencies during phoneme/word perception in audiovisual integration.

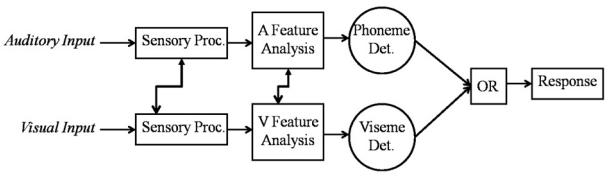


FIGURE 1 | A parallel model of audiovisual recognition with separate detectors for the auditory and visual channels. The model allows for cross-modal interaction in the early and later stages of recognition, although our study primarily focuses on obtaining evidence for later interactions.

Although later ERPs occurring post 400 ms have not been commonly analyzed in audiovisual word recognition, they are believed to be associated with phonemic recognition and semantic processing. In one a spoken word recognition study examining the relationship between semantic activation and ERP amplitude, Woodward et al. (1990) uncovered evidence for a large negative peak occurring around 480 ms, followed by a large positive potential peaking approximately 800 ms. The scalp tomography consisted of several frontal and parietal electrodes, and variability in latency and amplitude was believed to correspond to recognition points. Other studies on audiovisual integration have investigated late ERP components associated with conscious language recognition, although this is usually done within the context paradigms in which an "odd-ball" or incongruent signals are detected (cf. Klucharev et al., 2003; Riadh et al., 2004). Later potentials have also been shown to be influenced by the integration of incongruent audiovisual signals (e.g., Riadh et al., 2004; Arnal et al., 2011), and also important for processing phonological information (e.g., Klucharev et al., 2003; Henkin et al., 2009). The importance of analyzing later EEG activation cannot be overstated. In general, later ERP activation will be associated with accessing lexical memory, categorization, semantic processing, and even executive function. All of these functions are vital for language processing—especially under difficult conditions.

We therefore aim to investigate the relationship between audiovisual recognition and integration efficiency in greater detail. This study will establish a systematic relationship between capacity (a mathematical index of integration), and a late ERP related to language processing. We will examine audiovisual integration under easy listening conditions where the visual signal may be of little use, and under degraded listening conditions, where the visual information becomes increasingly helpful. The earlier N1 component will be examined as well.

MEASURING INTEGRATION EFFICIENCY

Integration efficiency can be measured by using a measure of *capacity* [$C(t)$]—a cumulative measure of work competed or energy expenditure (Townsend and Nozawa, 1995; Townsend and Wenger, 2004). It is a probabilistically defined RT measure in which independent first-terminating processing establishes a benchmark. Capacity is a measure that compares RTs from trials where auditory and visual information are present, to RTs

obtained from trials where either auditory-only or visual-only information is present. The capacity coefficient uses the entire distribution of RTs, at the level of the integrated hazard function. The integrated hazard function is defined as

$$H\left(t^*\right) = \int_0^{t^*} h\left(t\right) dt, \text{ where } h\left(t\right) = \frac{f\left(t\right)}{S\left(t\right)}$$

and *f(t)* is the probability density function and *S(t)* is the survivor function, such that *h(t)* gives the probability of a response in the next instant of time given that the response has not yet occurred (Townsend and Ashby, 1978, 1983; Townsend and Wenger, 2004). The hazard function approach has both conceptual and statistical advantages (see Wenger and Gibson, 2004 for discussion). Crucially, for our integration study, it captures the notion of capacity and "efficient energy expenditure" more closely than mean accuracy or RTs.

The use of capacity can thus be advantageous over mean RTs. First, as we shall see, capacity assays efficiency relative to independent (race) model predictions (Miller, 1982). Independent race models predict that auditory and visual information does not influence each other during processing; however, audiovisual processing is faster than either auditory or visual alone due to purely statistical reasons. Furthermore, context independence refers to the assumption that auditory completion times, for example, are unaffected by whether or not visual information is present (e.g., Townsend and Nozawa, 1995). Deviations from model predictions suggest that the predictions of the independent channels model have been falsified due to either limitations in efficiency or processing resources, facilitatory or inhibitory cross-channel interactions (e.g., Eidels et al., 2011), or perhaps coactivation where the auditory and visual information are pooled into a common processor (Miller, 1982; Townsend and Nozawa, 1995), in which case capacity is much greater than "1." A second advantage of capacity is that it makes use of integrated hazard functions. Given that the hazard function can be interpreted in terms of the instantaneous intensity of work, the integrated hazard function can be interpreted in terms of the total amount of work completed up until time *t*. Townsend and Nozawa (1995) derived the benchmark capacity coefficient for tasks in which observers are presented with 0, 1, or 2 target stimuli and have to respond if either 1 or 2 stimuli are present. For present purposes, if we let $H_{AV}(t)$ denote the integrated hazard function obtained from audiovisual trials, and let $H_A(t)$ and $H_V(t)$ denote the integrated hazard functions obtained from the auditory-only and visual-only trials, respectively, then the capacity coefficient is defined as:

$$C(t) = \frac{H_{AV}(t)}{H_A(t) + H_V(t)} \quad (1)$$

Note that the term in the denominator corresponds to the predictions of an independent race model (Miller, 1982). The capacity coefficient provides a useful non-parametric measure of integration efficiency in a variety of settings, with there being three possible outcomes and associated interpretations. First, the capacity coefficient can be greater than 1 at time *t*, indicating faster RT and thus more work completed in the audiovisual condition compared to the auditory- and visual-only conditions. In this case, we have highly efficient integration since RTs in the audiovisual condition are faster than would be predicted by independent race models. Second, capacity can be less than 1 at time *t*, indicating slower reaction times in the audiovisual condition compared to the unisensory conditions, and therefore inefficient audiovisual integration. A third possibility is that the capacity coefficient can be equal to 1 at time *t*. This would suggest that audiovisual processing is neither faster nor slower and is thus just as efficient as unisensory processing.

HYPOTHESES AND PREDICTIONS

We aim to model integration efficiency (i.e., Altieri, 2010; Altieri and Townsend, 2011) at different auditory S/N ratios in an audiovisual word identification task. We will relate neural measures of integration efficiency with behavioral measures of efficiency across variable S/N ratios. To accomplish this, we obtained EEG recordings and compared how peak and time-to-peak amplitudes in the audiovisual condition differed from the uni-sensory conditions as a function of auditory S/N ratio. For comparison purposes, we also report traditional accuracy based measures of integration ("Gain," e.g., Sumby and Pollack, 1954), although we would argue that accuracy alone does not reflect integration efficiency as meaningfully as capacity.

HYPOTHESES

ERPs

We aim to examine the hypothesis that the visual signal is used more (or less) efficiently as listening conditions change. Furthermore, the neural signal should co-vary with capacity observed in individual participants.

Null hypothesis

The null hypothesis for ERP data predicts that the relation between AV ERPs and A-only peak ERPs will remain constant across listening conditions. Of course, the amplitude of the signals should differ as a function of noise (likely decreasing in noisy listening conditions); however, the relative amplitude between AV and A should remain relatively constant. This should be true for the later ERP, and earlier N1 potentials.

Alternative hypothesis

We predict that the AV peak amplitude for the late ERP will increase relative the A-only (and possibly V-only) as listening conditions became increasingly degraded. This should mirror changes in capacity (discussed next). First, (1) in the high S/N ratio condition, the peak ERP occurring post 400 ms will be approximately equivalent in the multisensory and unisensory conditions; (2) in the −12 and −18 S/N ratio conditions, the amplitude will be greater in the AV compared to the unisensory conditions. This is because the AV ERP should increase as visual information increasingly assists auditory identification, the latter of which becomes degraded and requires visual place cues to facilitate recognition (Grant et al., 1998). Hence, as A-only accuracy, speed, and amplitude decrease, AV speed, accuracy and therefore amplitude should remain stable due to the presence of visual cues. This prediction is further motivated by evidence indicating

reductions in auditory ERP amplitude in patients with noise induced hearing loss due to tinnitus (Attias et al., 1993) and in normal-hearing participants as noise thresholds change (Martin and Stapells, 2005; see also Woodward et al., 1990; Stevenson et al., 2012). Martin and Stapells (2005) observed that increased noise delivered via low pass filtering reduced auditory N1, N2, and P3 amplitudes, and also time-to-peak. Together, we predict that complementary cues provided by the visual signal in the AV condition should enhance recognition to a greater degree under lower S/N ratios.

Lip-movement typically precedes the auditory signal by tens of milliseconds in ecologically valid speech signals. Researchers have also argued that the leading visual speech cues provide predictive information that modulates early auditory encoding (e.g., van Wassenhove et al., 2005); effects of visual lead have been shown to facilitate auditory encoding, which is reflected in amplitude changes in the N1-P2 complex. We thus predicted that the N1 ERP amplitude associated with visual prediction would be greater for auditory-only stimuli vs. audiovisual stimuli in the high S/N ratio condition (e.g., Besle et al., 2004; van Wassenhove et al., 2005; Pilling, 2009). We also predicted that this difference between the audiovisual and auditory-only ERPs may be attenuated for lower auditory S/N ratios in which capacity increases.

CAPACITY

The *null hypothesis* for capacity likewise predicts that integration will not change as a function of auditory S/N ratio within an individual listener. Incidentally, accuracy based models of integration often predict that each individual has a certain pre-established integration ability that does not change across listening conditions, contexts, or environments (Braida, 1991; Grant et al., 1998). To use one example, the Fuzzy Logical Model of Perception (FLMP; Massaro, 2004) predicts optimal integration of auditory and visual speech cues regardless of the perceived quality of the auditory and visual signals. This concept of optimality can perhaps best be translated in the capacity approach by assuming that optimal integration implies unlimited (or even super) capacity.

Our *alternative hypothesis* mirrors ERP hypotheses by predicting that capacity will be inefficient [$C(t) < 1$] for high S/N ratios (clear signal), but become efficient [$C(t) > 1$] for lower S/N ratios (-12 to -18 dB). Capacity should be limited in ideal listening environments since normal-hearing listeners do not normally utilize visual speech cues in such conditions. This is manifested in RTs by virtue of the fact that the AV distribution of RTs should not be much different than the auditory-only one (see Altieri and Townsend, 2011). Of course, as the auditory-only becomes slower in degraded conditions, the AV RT distribution becomes faster relative to the unisensory ones. The ERPs mirror capacity predictions because multisensory ERPs should fail to show evidence for visual gain (AV > A-only) in the clear listening condition. Hence, the EEG signal in the multisensory condition should not be sufficiently better than the one evoked by the auditory-only condition. These predictions are motivated by the law *of inverse effectiveness*, which stipulates that as auditory and visual-only recognition become less "effective," AV integration improves relative to unisensory recognition speed/accuracy (e.g., Stein and Meredith, 1993; Stevenson et al., 2012). Likewise, cross-modal stochastic resonance (SR), similar to inverse effectiveness, predicts that the addition of noise to unisensory signals facilitates the detection of multisensory signals. However, SR differs from inverse effectiveness because it assumes that there is an optimal level of noise that can be added to a signal to achieve the maximum multisensory benefit (Ross et al., 2007; Liu et al., 2013).

METHODS

PARTICIPANTS

Four young (age range of 20–28) right-handed native speakers of American English (1 female) were recruited from the student population at The University of Oklahoma. Participants reported normal or corrected to normal vision, and no participant reported having neurological or hearing problems. Participants were paid \$8/h for their participation[1]. The Institutional Review Board at The University of Oklahoma approved this study.

This study obtained a sufficient number of data points to adequately estimate integrated hazard functions to compute robust capacity measures (Townsend and Nozawa, 1995), while also providing sufficient power to compare ERPs across conditions for each individual. Capacity and ERPs are time variable measures capable of showing differences in performance at different time points. Both capacity scores and analyses showing differences in ERPs will be displayed for each individual. Capacity also functions as a diagnostic tool for capturing information processing strategies at the level of the individual (e.g., Townsend and Nozawa, 1995; Townsend and Wenger, 2004; Townsend and Eidels, 2011; see Estes, 1956, for problems with averaging data). Our strategy should prove exceedingly useful for diagnosing audiovisual integration skills that can ostensibly vary as a function of auditory clarity, cognitive workload, or audiometric configuration, even within one individual (e.g., Altieri, 2010; Altieri and Townsend, 2011).

STIMULI

The stimulus materials consisted of audiovisual full-face movie clips of two different 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 a study carried out by Altieri (2010) and Altieri and Townsend (2011). The auditory, visual, and audiovisual movies were edited using Final Cut Pro HD 4.5. Each of the auditory files was normalized during the digitization process and sampled at a rate of 48 kHz (16 bits).

[1]These same subjects participated in a non-speech pilot study for three blocks of 800 trials over a period of 3 days. The study involved presenting visual stimuli (Gabor patches), auditory pure tones presented at three auditory S/N ratios: clear, -12 dB, and -18 dB), and simultaneously presented (AV) Gabor patches and auditory pure tones, in addition to catch trials consisting of white noise and a blank screen. Participants were required to make a "yes" response by pressing the right button on the mouse if a Gabor patch appeared on the screen, they heard an auditory tone, or saw both a Gabor patch and auditory tone. They were required to respond "no" by pressing the left mouse button on blank catch trials. As in the primary experiment, processing capacity was computed for each auditory S/N ratio, and EEG recordings were obtained via a dense electrode 128-channel net.

Each movie was digitized and rendered into a 720 × 480 pixel clip at a rate of 30 frames per second. Video stimuli were played with a refresh rate of 60 Hz. The duration of the auditory, visual, and audiovisual files ranged from 800 to 1000 ms. White noise was mixed with each auditory file using Adobe Audition. This allowed for the creation of S/N ratios of −12 dB and −18 dB, in addition to a clear auditory S/N ratio in which noise was not mixed in with the stimuli.

The eight words in the stimulus set were presented in a total of seven blocks, including an AV-clear, AV−12, AV−18, A-clear, A−12, A−18, and V-only block. Each block consisted of 240 total trials, including 120 trials spoken by each talker. Each of the 8 words was presented a total of 30 times per block (15 spoken by each talker). In total, the experiment consisted of 1680 trials distributed over seven sessions within one 2-week period.

While the inclusion of a limited response set size was important for obtaining accurate RTs across a large number of trials and conditions, a potential disadvantage to this approach is that a closed stimulus set of 8 words lacks a degree of ecological validity. Listeners may process words differently compared to real world settings. For example, lip-reading accuracy scores will be higher for a set size of 8-monosyllabic words compared to a larger response set (Sumby and Pollack, 1954), or a sentence processing task (Altieri et al., 2011). One may object that the small set size encouraged listeners to recognize stimuli by relying on simple pattern recognition rather than word recognition. We remedied this by requiring participants to respond by pressing a button corresponding to the word they thought the talker said. The intent was to encourage listeners to engage in word recognition. This is in contrast to previous approaches which have required binary responses from participants to syllables (e.g., Massaro, 2004) or words (Winneke and Phillips, 2011). More importantly, if the words in our study were processed using pattern recognition based on simple auditory and visual features, it should be reflected in the capacity analysis. A preponderance of studies assessing the race model inequality using simple auditory or visual features, such as tones and dots (e.g., Miller, 1982; Berryhill et al., 2007) have consistently shown upper bound violations on independent race model predictions. When the upper bound on processing speed is violated, it indicates the presence of cross-modal dependencies and hence, a violation of independence. As discussed later, our pilot study using Gabor patches and auditory pure tones showed similar evidence for super capacity (as fast RTs) across each S/N ratio. This reflects a radically different profile from the capacity data in the word recognition experiment. Hence, the divergence in capacity results between simple auditory-visual detection and word recognition experiments indicates vastly different processing strategies—namely deeper processing for linguistic stimuli.

As a final caveat, noise was premixed with the stimuli prior to the experiment. Research indicates that participants may learn meaningless noise sounds over the course of many trials (e.g., Agus et al., 2010). However, our randomized block design, and the fact that each participant exhibited low accuracy scores in the low S/N ratio conditions (see Results), indicates that significant learning of noise patterns did not occur. Finally, while white noise may lack the properties of other masking strategies such as multi-talker babble that are most appropriate for sentence length materials (e.g., Bent et al., 2009), it still significantly reduces performance on vowel and consonant intelligibility (Erber, 2003).

EEG RECORDING

EEG recordings were made using EGI NetStation system with a 128-channel electrode net (Electro Geodesics International, Eugene, OR). Data were acquired continuously throughout the session and sampled at a rate of 1 kHz. The electrodes were referenced to the central (Cz) electrode. A significant advantage of using Cz as a reference electrode is that it is centrally located, and provides a reference that equal distances between electrodes on each hemispheres. The purpose was to obtain a central head location from which each frontal and parietal electrode could be referenced. Two electrodes, one located under each eye monitored eye movements, and a set of electrodes placed near the jaw were used for off-line artifact rejection. Channel impedances were maintained at 50 K Ohms or less for the entire testing session.

After down-sampling the data to 250 Hz, bad channels were identified and eliminated by visual inspection and ocular and other artifacts were removed automatically using EEGLAB V. 9 (http://sccn.ucsd.edu/eeglab/) with a statistical thresholding technique for detecting significant deviations. Baseline correction was carried out using an interval of 400 ms prior to the onset of the stimulus (i.e., word) in each condition (AV, A-only, and V-only). Data were organized into seven categories according to stimulus condition: AV (clear signal), AV (S/N ratio = −12 dB), Audiovisual (S/N ratio = −18 dB), A-only (clear signal), A-only (S/N ratio = −12 dB), A-only (S/N ratio = −18 dB), and V-only. The overall proportion of trials not rejected due to noise or artifacts per condition was over 0.90 for each condition [0.98 (AV clear), 0.94 (AV −12 dB), 93 (AV −18 dB), 0.98 (A clear), 0.93 (A −12 dB), 0.91 (A −18 dB), and 0.94 (V-only)]. Individual averages were computed at each time point for each electrode, with these averages computed for correct responses. All data were low-pass filtered at 55 Hz. A total of 36 electrodes (18 located on the frontal scalp region, and 10 located in the left parietal, and 8 in the left temporal scalp regions) were included in the data analysis. We selected a montage that included electrodes analyzed in previous studies, including left FC, C3, and CP (Pilling, 2009).

ERPs and times-to-peak-amplitudes for and participant were computed by obtaining the values of minima and maxima within specific time windows following stimulus onset. The primary peak ERP component of interest was the peak corresponding to phonological/language processing occurring roughly 400–700 ms post stimulus. Sometimes these peaks have been reported as being negative (depending on electrode positioning), although positive peaks connected to auditory language processing have been observed (e.g., Henkin et al., 2009; see Mehta et al., 2009, for discussion on the "P6" in word recognition). For the later ERP, we used the interval from 400 to 700 ms. We calculated positive peak amplitude values within this window that were significantly greater than 0, and the time to that peak using a maximum peak detection algorithm. For the N1 potential, we computed the minimum value in the trough (and time to negative peak) occurring

between 70 and 120 ms post stimulus. The mean ERP value was calculated and submitted for analysis when the peak value for a given component differed significantly from 0 in an electrode.

PROCEDURE

Participants were seated at a fixed distance of 76 cm in front of a black and white CRT computer monitor with their chin placed on a chin rest. Experimental stimuli were presented using E-Prime version 2.0, and interfaced with NetStation software (EGI, Eugene OR) for the collection of continuous EEG recordings. Auditory stimuli were played via two speakers situated approximately 60 cm to the side.

Experimental trials began with a fixation cross (+) appearing in the center of the monitor followed after 200 ms by the stimulus. The stimuli were either auditory-only, visual-only or audiovisual trials, with each of these trials presented in separate blocks. Auditory stimuli were played at a comfortable listening volume (approximately 68 dB). Responses were collected via button press using the computer keyboard. Each of the buttons (1–8) was arranged linearly on the keyboard and was labeled with a word from the stimulus set. The labeling configuration was controlled across participants. 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 (from the time that the previous trial terminated once a response was detected). 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. Auditory-only trials were played with a blank computer screen. Likewise, visual-only trials were played without any sound coming from the speakers. The screen went blank once each trial containing a video was terminated. Each session consisted of one randomly ordered block per day and lasted approximately 45 min. To avoid order effects, the experimental blocks were randomized and presented in a unique order for each participant. Participants received 48 practice trials at the onset of each experimental block; data from these trials were not included in the subsequent analyses. Participants learned the keyboard response mappings during these practice trials such that head and eye movements were kept to a minimum.

RESULTS

BEHAVIORAL ANALYSES

The behavioral data were analyzed at two levels. First, mean accuracy and RT were examined across participants and auditory S/N ratios. This allowed for a coarse assessment of changes in integration efficiency as a function of S/N ratio. This method is less sensitive to fine gained temporal changes in efficiency relative to the analyses performed at the level of the distributions using the capacity coefficient (Townsend and Ashby, 1978, 1983; Wenger and Gibson, 2004).

Table 1 displays mean accuracy and RT results for each of the four participants, in addition to the mean and standard deviation (SD) across participants. The terms AV, A, and V denote the mean accuracy scores for the audiovisual, auditory, and visual conditions, respectively, while AV (RT), A (RT), and V (RT) denote the mean (SD) RTs in each of those conditions. Visual "Gain" (e.g., Sumby and Pollack, 1954; see also Grant, 2002; Bergeson and Pisoni, 2004) quantifies the relative benefit or the participants receives (in accuracy) by having the visual signal present in addition to the auditory signal. That is, what is the proportional gain in accuracy achieved by being able to see a talker's face? This is estimated as: $Gain = [AV - A]/[1-A]$; higher numbers indicate more efficient use of visual information, with 1 being the highest possible gain. In cases of extremely high unimodal accuracy, gain scores may become difficult to interpret. For example, Participant 4 showed a gain of −1.0, which results from a slightly lower AV relative to A-only score. However, both scores are effectively near ceiling, making the gain score of −1.00 meaningless in this case (in actuality, the data show an absence of gain). Visual gain in the temporal domain, labeled "Gain (RT)," signifies the overall benefit received in the RT domain from the presence of the visual signal. It is estimated as $Gain\ (RT) = A(RT) - AV(RT)$. The proportion of auditory gain (gain afforded by the auditory speech signal over and above the visual) is also provided in the table $Gain\_A = [AV - V]/[1-V]$, as is the RT analogue $Gain\_A(RT) = V(RT) - AV(RT)$.

Results from the clear auditory condition are shown in **Table 1A**, the −12 dB condition in **Table 1B**, and the −18 dB in **1C**. On average, identification accuracy in the visual-only condition was 75% with the three out of four participants scoring ∼70% and one scoring 90%. This observation was consistent with previous findings in an 8-alternative forced-choice task (Sumby and Pollack, 1954; Altieri and Townsend, 2011).

The results in **Table 1A** reveal virtually no difference in accuracy between the AV and A condition across subjects. Not surprisingly, Gain scores were close to 0 for each participant in the clear condition. The RT analyses revealed little difference between audiovisual and auditory trials; RT Gain scores were near zero, revealing that the visual signal failed to facilitate processing in the temporal domain. Overall, the RT results suggest that audiovisual integration either did not occur in this condition, or possibly that it either did not provide any benefit or extract any cost.

In the −12 dB S/N condition (**Table 1B**), recognition accuracy in the audiovisual condition was higher than in the auditory-only condition. Gain scores for each participant were approximately 70% or greater, with an overall mean of 75%. Similarly, a noticeable visual gain was observed in the RT data, with AV RTs being nearly 700 ms faster on average compared to auditory-only RTs. This level of gain was statistically greater than that observed in the clear condition [$t_{(3)} = 3.9$, $p < 0.05$].

The −18 dB S/N ratio condition (**1C**) revealed a pattern of results similar to the −12 dB S/N ratio. Auditory-only recognition accuracy was considerably above chance for each participant although performance in all of the conditions was extremely degraded. Nonetheless, audiovisual recognition accuracy (mean 92%) was markedly higher compared to auditory-only accuracy (mean 33%). Consequently, proportional Gain scores were significantly higher in the −18 compared to the −12 dB condition [$t_{(3)} = 4.3$, $p < 0.05$]. Interestingly, accuracy scores in the audiovisual, auditory-only, and visual-only conditions were consistent with those observed in previous word identification studies

Table 1 | This table displays the mean accuracy scores for the audiovisual (AV), auditory-only (A), and visual conditions (V).

	Sub. 1	Sub. 2	Sub. 3	Sub. 4	Mean	SD
(A) RESULTS FOR THE *CLEAR* AUDITORY SIGNAL CONDITION						
AV	0.98	0.98	0.97	0.98	0.98	0.01
A-Only	0.98	0.98	0.97	0.99	0.98	0.01
V-Only	0.71	0.67	0.90	0.69	0.75	0.11
Gain	0.00	0.00	0.00	−1.00	−0.25	0.50
Gain_A	0.93	0.94	0.67	0.94	0.87	0.13
AV (RT)	1455 (310)	1586 (286)	1273 (324)	1272 (413)	1397	153
A (RT)	1466 (585)	1583 (456)	1253 (267)	1280 (255)	1396	157
V (RT)	1705 (464)	1946 (458)	1405 (291)	1771 (472)	1707	225
Gain (RT)	11	−3	−20	8	−1	12
Gain_A (RT)	250	360	132	499	310	157
(B) RESULTS FOR THE −*12 dB* AUDITORY CONDITION						
AV	0.93	0.93	0.90	0.95	0.93	0.02
A-Only	0.72	0.73	0.69	0.69	0.71	0.02
V-Only	0.71	0.67	0.90	0.69	0.75	0.11
Gain	0.75	0.74	0.68	0.84	0.75	0.07
Gain_A	0.76	0.79	0	0.84	0.60	0.40
AV (RT)	1263 (344)	1174 (521)	1361 (341)	1296 (244)	1274	78
A (RT)	1966 (778)	2271 (572)	1604 (464)	1958 (625)	1950	273
V (RT)	1705 (464)	1946 (458)	1405 (291)	1771 (472)	1707	225
Gain (RT)	703	1097	243	662	676	349
Gain_A (RT)	442	772	44	475	433	299
(C) RESULTS FOR THE −*18 dB* AUDITORY CONDITION						
AV	0.94	0.90	0.95	0.89	0.92	0.03
A-Only	0.33	0.33	0.30	0.36	0.33	0.02
V-Only	0.71	0.67	0.90	0.69	0.75	0.11
Gain	0.91	0.85	0.93	0.83	0.88	0.05
Gain_A	0.79	0.70	0.50	0.65	0.66	0.12
AV (RT)	1365 (253)	1708 (199)	1331 (269)	1384 (345)	1447	175
A (RT)	2223 (786)	2748 (795)	1938 (605)	2076 (577)	2246	354
V (RT)	1705 (464)	1946 (458)	1405 (291)	1771 (472)	1707	225
Gain (RT)	859	1040	607	692	800	191
Gain_A (RT)	340	238	74	387	260	139

The Gain scores {[AV − A]/[1 − A] & [AV − V]/[1 − V]} and RT Gain scores {(ART − AVRT) & (VRT − AVRT)} are shown as well.

using 8-alternative forced-choice tasks and identical S/N ratios (e.g., Sumby and Pollack, 1954). Visual gain in the RT data, under the most degraded listening conditions, was also significantly greater compared to the gain scores in clear listening condition [$t_{(3)} = 28$, $p < 0.001$], but not compared to the −12 dB condition. Taken together, mean accuracy and RT results indicate that the most efficient integration occurs between −12 and −18 dB, and that integration may not need to occur when listening conditions become less degraded due to ceiling effects in the auditory modality (see Ross et al., 2007).

Capacity and integration efficiency

Figure 2A shows example cumulative distribution functions (Participant 4) obtained from the audiovisual, auditory, and visual-only conditions. Results are shown across each S/N ratio. The Miller bound [$F_A(t) + F_V(t)$] was violated across several time points in the lower S/N ratio conditions. Interestingly, $F_{AV}(t)$ was less than $F_A(t)$ (the fastest unimodal condition) across most time points, indicating lower Grice bound (Grice et al., 1984) violations (Grice bound = max{$F_A(t)$, $F_V(t)$}). The Grice bound sets a lower limit on independent processing, where violations suggest negative inter-modal dependencies, and inefficient integration.

The capacity coefficient, *C(t)*, was calculated for each participant across the three listening conditions (clear, S/N = −12 dB, S/N = −18 dB). Capacity function values are plotted as symbols in **Figure 2B** (correct responses were used in the capacity calculations). Capacity analyses were computed by pooling RT data across the 8 words in the computation of the cumulative hazard functions shown in Equation 1. While pooling data across stimuli with different onset consonants (e.g., "b" vs. "sh") may obscure effects for individual words, the same overall trend was observed across each stimulus (see **Appendix**). A greater audiovisual benefit, in terms of both mean RT and accuracy, was observed for each stimulus in the −12 and −18 dB conditions.

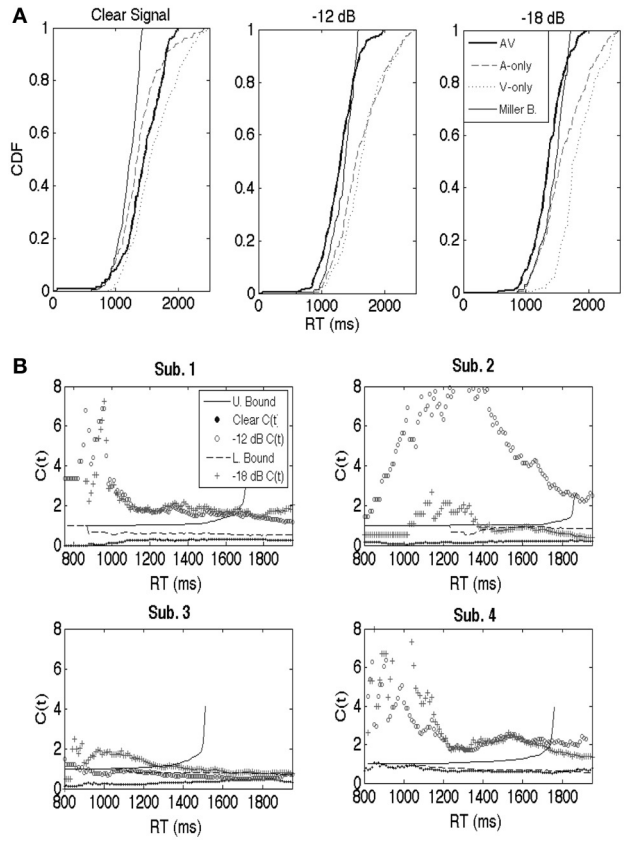


FIGURE 2 | (A) For purposes of illustration, cumulative distribution functions (CDFs; participant 4) are shown separately for the auditory-only, visual-only, and audiovisual trials for the clear, −12 dB, and −18 dB listening conditions. Miller and Grice bounds are plotted as well. **(B)** Similar to the CDFs, capacity is a continuous function. Results are shown separately for participants 1 through 4. Participant 1's results are shown in the upper left, 2's in the upper right, 3's in the lower left, and 4's in the bottom right. Each panel displays *C(t)* across each listening condition, as well as the upper (Miller) and lower (Grice) bounds translated into capacity space (cf. Townsend and Eidels, 2011 for equations). The symbols (clear, −12 dB, and −18 dB) denote the values for the capacity function. Note that there should be separate upper and lower for each auditory S/N ratio. Our strategy was to plot the averaged bounds to avoid clutter, although interpretations of the data would not have changed regardless of the bounds used.

Hence, mean RTs were considerably faster, and mean accuracy was also greater in the audiovisual condition compared to either the auditory or visual-only conditions. Conversely, none of the stimuli showed evidence consistent with an audiovisual benefit in the "clear" condition, just as expected.

The capacity results in **Figure 2B** followed a similar pattern across participants: limited capacity in the "clear" condition, and efficient integration marked by violations of the Miller bound, at least at some time points, in more difficult listening conditions. The upper or Miller Bound is depicted by the solid line and represents an upper limit on independent race model predictions. Violations of these bounds in the −12 and −18 dB conditions strongly suggests violations of independence, and hence, facilitatory cross-modal dependencies. The *C(t)* results for the clear listening condition hover well below 1 and near 1/2 across nearly all time points and participants. Consequently, *C(t)* violated the lower bound in every participant for at least a few time points. The All of this serves to clarify the ambiguity with respect to integration obtained from the mean data. Recall that those data suggested either inefficient on non-existent integration. The capacity coefficients clearly show that the integration was in fact extremely inefficient. The lower bound represents a lower limit on independent race model predictions and is represented by the dashed line in each panel[2]. The *C(t)* data for each participant in the −12 dB and −18 dB S/N ratios showed consistent violations of the upper bound, particularly for early response times. Although the results revealed individual differences (i.e., lower efficiency for Participant 3, and higher capacity in the −12 dB than the −18 dB condition for Participant 2), the qualitative pattern of results held across participants.

Thus, the results show rather strong evidence in favor of the predicted pattern: inefficient audiovisual integration under optimal listening conditions but highly efficient integration under degraded listening conditions. The ubiquitous violations of the upper and lower bound strongly suggests facilitatory interactions in the case of upper bound violations, and inhibitory interactions in the case of lower bound violations (e.g., Eidels et al., 2011). As shown in **Figure 1**, interactive models with separate decisions on the auditory and visual modalities can account for such violations via interactive mechanisms across channels that change from inhibitory to excitatory as a function of the clarity of the auditory signal. Such an account is consistent with the idea that extensive uni-sensory processing takes place in auditory and visual pathways, and that interactions occur even in the later stages of recognition (Bernstein et al., 2004; Ponton et al., 2009; Naue et al., 2011).

ERP ANALYSIS

Late peak

Figure 3 displays averaged ERPs calculated across electrodes from the frontal and parietal/temporal scalp regions for purposes of illustration. Results are shown from audiovisual (AV) auditory-only (A) and visual-only (V) ERPs across three S/N ratios. ERPs were smoothed using a moving window approximation. ERP amplitudes were determined within an electrode by utilizing a function that computed the maximum peak value within the time window. First, we utilized a time window ranging from 400 to 700 ms when determining the peak value, and also the latency at which it occurred. We can observe that peaks emerged in the audiovisual condition, on average, in the 400-500 ms time window. While the late peak was generally reduced in the low S/N auditory-only conditions, significant potentials did emerge post 400 ms, highlighting the importance of analyzing later potentials in language perception studies. A positive visual evoked potential (∼200 ms) was also observable across frontal electrodes, and a negative potential due to a polarity reversal was observed in temporal/parietal electrodes. The auditory and audiovisual amplitudes were generally positive across anterior and posterior scalp

[2]Townsend and Eidels (2011) translated the upper Miller bound, and lower Grice bound into a unified capacity space. **Figure 2B** depicts these bounds in each of the four panels.

regions, although one may observe that the auditory potentials were considerably attenuated, even to the point of becoming slightly negative for later times, as noise increased. The quantification procedure of using positive peak was consistent for both frontal and temporal/parietal sites.

For the initial statistical analysis on the late amplitude, we carried out a One-Way ANOVA on individual electrodes across all participants to determine whether there was a main effect for modality (α levels were 0.05 unless otherwise indicated). The ANOVA on the aggregate data indicated a significant effect for modality. **Figure 4A** plots results across the frontal and left parietal/temporal regions (AV − A) for the ERP as a function of audiovisual gain. All error bars represent one standard error of the mean peak amplitude calculated across individual electrodes. In order to illustrate the quantitative relationship between capacity values and ERPs, **Figure 4B** displays the ERP enhancement scores as a function of capacity obtained in the clear, −12 dB, and −18 dB listening conditions, respectively. Maximum capacity values were obtained for each auditory S/N ratio across the intervals displayed in **Figure 2** and averaged across individual observers (e.g., Altieri and Townsend, 2011). **Figure 4** shows that as audiovisual gain and capacity increased the difference between the audiovisual and the unisensory signals also increased.

In both frontal and left regions, visual speech significantly enhanced the late ERP amplitude [$F_{(2,\ 629)} = 3.3,\ p < 0.05$] compared to the auditory and visual-only ERPs. This significant differences in multisensory vs. unisensory ERPs suggests that changes in amplitude were not merely the superposition of component effects (in which the AV peak amplitude would equal the sum of the A and V-only peaks, AV = A + V). The ANOVA testing the interaction between region and modality was significant, indicating that the strongest effects occurred in frontal regions compared to left parietal/temporal regions [$F_{(4,\ 629)} = 2.8,\ p < 0.05$]. The interaction can be observed in **Figures 4A,B**, which shows greater ERP amplitude increase the frontal region compared to the left regions. The observed interaction, and the fact that the changes in amplitude were not merely the superposition of component effects (where the AV peak amplitudes simply reflect the sum of the auditory-only and visual-only peak amplitudes, AV = A + V). The ERP analysis also evidenced significant enhancement compared to the auditory-only (AV vs. A) ERP peak [$t_{(629)} = 2.2, p < 0.05$]. These findings appear contrary to previous literature indicating that the presence of visual speech in the AV condition should yield a reduction rather than enhancement in peak amplitude (e.g., van Wassenhove et al., 2005; Pilling, 2009; Winneke and Phillips, 2011).

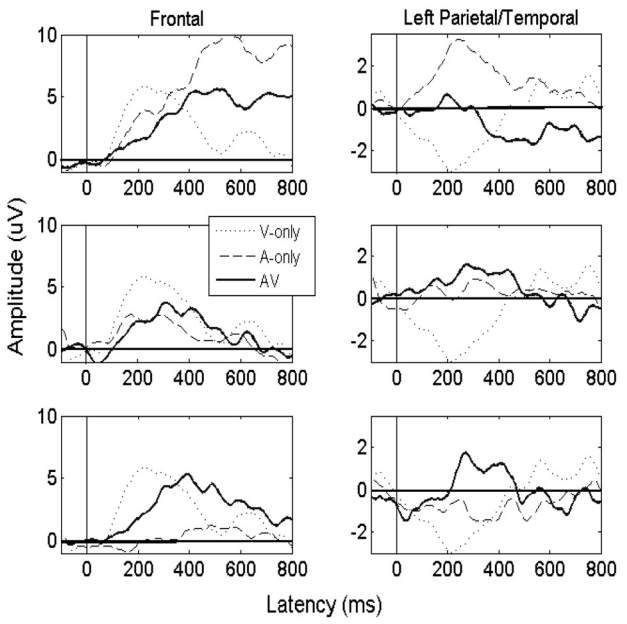


FIGURE 3 | Figure displaying averaged ERPs obtained from the "clear" auditory condition (top), the −12 dB S/N ratio (middle), and the −18 dB S/N ratio (bottom). The solid line shows the averaged ERPs for the audiovisual (AV) condition, and the dashed line represents the auditory-only (A) condition. The dotted line represents the visual-only (V) condition, which is strongly expressed in frontal regions due to feed-forward connections originating from occipital brain regions. Results are shown using sample electrodes from the frontal scalp region (e.g., F3, F7, and Fz), and Parietal/Temporal Region (e.g., C3, P3, and T5).

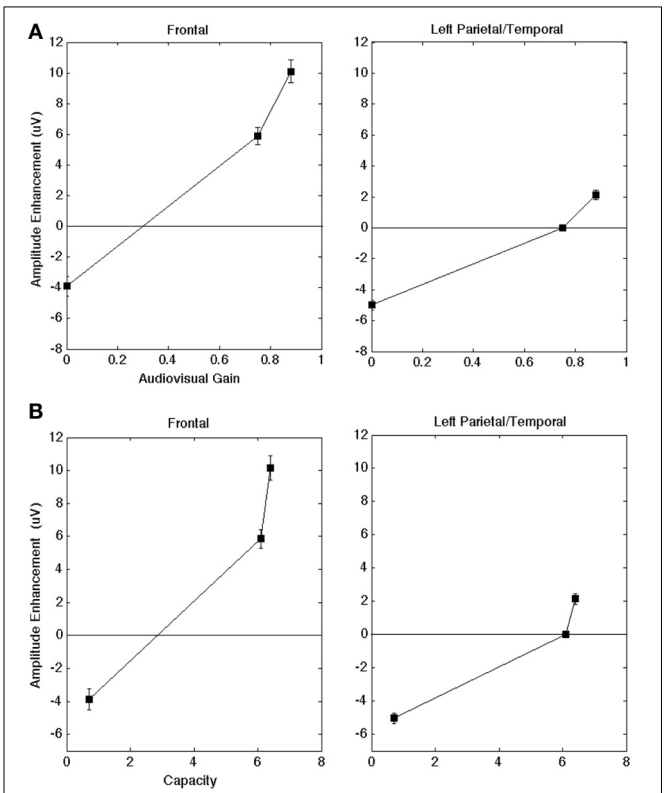


FIGURE 4 | (A) AV gain in amplitude for the peak ERP (AV − A) as a function of audiovisual gain across brain regions of interest. A positive value means that the average AV amplitude was larger than the A. The scores are collapsed across each of the four participants. **(B)** Audiovisual gain in amplitude as a function of capacity scores (in the clear, −12 dB, and −18 dB S/N ratio conditions, respectively) across brain regions of interest. The scores are collapsed across each of the four participants. Error bars denote one standard error of the mean computed across individual electrodes (across subjects) within a given region.

The reason for the observed discrepancies likely lies in the fact that audiovisual integration mechanisms operate differently across listening conditions. Previous studies [e.g., van Wassenhove et al. (2005) and Pilling (2009)] analyzed the N1/P2 ERP components under clear auditory listening conditions. Interestingly, the contrasts for the different S/N ratio conditions support this hypothesis. The AV_{High} vs. A_{High} contrast in the mean ERP (clear condition) yielded the predicted reduction in the audiovisual peak amplitude [$t_{(631)} = -3.2$, $p = 0.001$]. Next, the contrast for the AV_{Low} vs. A_{Low} showed strong evidence for AV enhancement (i.e., AV > A) [$t_{(629)} = 5.1$, $p < 0.001$], although the AV_{Med} vs. A_{Med}only showed evidence for a non-significant trend ($p = 0.11$) toward AV enhancement.

The results for the *t*-test for each of the four individual participants are shown in **Table 2**[3]. The key analyses for each participant included *t*-tests assessing the overall AV − A contrast on peak amplitude (across all frontal and parietal/temporal electrodes for the observer), and the contrasts for the high, $A_{High}V - A_{High}$, medium $A_{Med}V - A_{Med}$, and low $A_{Low}V - A_{Low}$ S/N ratio experimental conditions. Participants 2, 3, and 4 showed evidence for audiovisual enhancement (AV − A >0). Participant 1's results diverged from the other 3 participants in that an overall audiovisual reduction rather than enhancement was observed in the lowest S/N ratio listening condition. Frontal regions showed a significant reduction in the −18 dB condition. However, in the left parietal/temporal scalp regions, reduction was observed in the high S/N ratio while enhancement was observed in the −18 dB condition {the interaction between region and condition was significant [$F_{(7,\ 144)} = 14.1$, $p < 0.001$]}.

N1 component

We now briefly summarize data from the N1 component to bolster claims showing evidence for early audiovisual interactions during encoding (e.g., Besle et al., 2004; van Wassenhove et al., 2005; Pilling, 2009; Stevenson et al., 2012). A small negative amplitude (N1 ∼70–120 ms) was observed in the audiovisual conditions, and sometimes in the auditory-only. One reason why the early AV amplitude may have been similar to the A-only is that the visual signal essentially failed to provide useful bottom-up sensory information (although cf. van Wassenhove et al., 2005). However, under the medium (−12 dB) listening conditions, the visual signal likely provided early bottom-up sensory input that could eventually be combined with the degraded auditory signal. Interestingly, the results suggest that the N1 amplitude of the AV signal was once again reduced relative to the A-only in the −18 dB condition. Our preliminary explanation is that when the auditory signal became sufficiently degraded, the visual signal once again failed to provide sufficiently bottom-up sensory support. Nonetheless, as processing progressed, auditory phonemic information could be effectively extracted and integrated with visual cues (as observed by increased capacity and enhancement of the later ERPs). Of course, there exist S/N ratios in which the auditory signal becomes so degraded that the visual signal fails to be of any benefit (see Sumby and Pollack, 1954; Ross et al., 2007).

[3]The α level was set to the more conservative level of 0.01 to adjust for multiple comparisons.

For the statistical analyses, the ANOVA testing the interaction between region and modality was significant [$F_{(4,\ 508)} = 14.1$, $p < 0.001$]. This indicates that a greater negative peak amplitude (in the AV vs. A-only) occurred in the frontal compared to the left regions, mainly in the −12 dB condition. Individual contrasts for the N1 are also shown in **Table 2**. These data point to multisensory enhancement for Participants 1 through 4 in the −12 dB S/N (and an overall enhancement in Participants 1, 2, and 4 driven by the −12 dB condition). Although results diverged from previous findings showing AV suppression, our ERP results are in agreement with previous literature showing that the visual signal interacts with the auditory neural processing during early attentional and encoding stages. The difference in our task and previous studies employing discrimination with short matched/mismatched consonants (e.g., van Wassenhove et al., 2005) may help account for observed differences in early components.

Time-to-peak analysis

The time-to-peak analyses were less consistent across participants, but they still provided intriguing insights. Once again, we carried out one-way ANOVAs ($\alpha = 0.05$) using data obtained from individual electrodes across participants. The results from the combined data analysis on the late ERP for time-to-peak-amplitude demonstrated significant effects for modality. First, the presence of visual speech contributed to an overall slowdown in the time-to-peak for the late ERP [$F_{(2,\ 539)} = 4.9$, $p < 0.01$].

Table 2 | This table displays contrast results for the ERP peak amplitude for Participants 1 through 4.

Contrast	N1	Late ERP
SUB. 1		
AV − A	2.30*	−1.03
High	1.10	1.17
Medium	2.70**	−0.21
Low	0.88	−4.10***
SUB. 2		
AV − A	2.40*	0.49
High	0.41	−4.20**
Medium	3.20**	0.70
Low	0.80	3.60***
SUB. 3		
AV − A	1.90	4.40***
High	0.89	2.70**
Medium	3.60***	0.31
Low	1.50	3.60***
SUB. 4		
AV − A	2.45**	0.50
High	0.90	−4.20***
Medium	3.90***	0.70
Low	1.10	3.60***

*Signed numerical t-values and significance (***$p < 0.001$; **$p < 0.01$; and *$p < 0.05$) are shown in each cell corresponding to the high, medium, and low listening condition.*

The t-tests for difference of means from the AV − A contrasts ($\alpha = 0.01$) showed evidence for slower audiovisual processing when the auditory S/N ratio was clear/high [$t_{(558)} = 5.8$, $p < 0.001$], but facilitation when the auditory S/N ratio was −18 dB [$t_{(558)} = -4.3$, $p < 0.001$]. Interestingly, an examination of the $AV_{Time} - A_{Time}$ contrasts across listening conditions showed evidence for a slowdown in AV time-to-peak in the clear listening condition [$t_{(339)} = 2.6$, $p = 0.01$], and a trend in the same direction for the −12 dB S/N ratio [$t_{(339)} = 2.3$, $p = 0.02$]. **Figure 5** shows the temporal facilitation effects for the frontal and left parietal/temporal regions ($AV_{Time} - A_{Time}$) as a function of reaction time (RT) gain.

The time-to-peak contrasts ($AV_{Time} - A_{Time}$) for each participant are shown in **Table 3**. First, Participant 1 showed evidence for audiovisual temporal slow-down in the time-to-peak measurement in the clear listening condition, although the data for Participants 2 and 4 showed evidence for facilitation. Conversely, Participants 2, 3, and 4 showed evidence for temporal slow-down in either the −12 or −18 dB conditions. This analysis broken down by individual subjects data supports to the hypothesis that cross-modal interactions occur in the later stages of integration as phonetic and word recognition unfold (e.g., van Wassenhove et al., 2005; Ponton et al., 2009).

In summary, the accuracy, capacity, and ERP results provide converging evidence that poorer listening conditions afford the greatest efficiency in audiovisual integration. These results suggest that visual information influenced neural integration processes and were responsible for the observed effects on ERP peak amplitudes.

GENERAL DISCUSSION

The purpose of this study was to assess integration efficiency under different listening conditions while investigating how efficiency relates to brain activity. We proposed that capacity represents a continuous measure of efficiency (e.g., Townsend and Nozawa, 1995; Townsend and Wenger, 2004). This approach assumes that word recognition occurs as soon as either auditory or visual information (corresponding to a specific word/category) reaches a threshold (see Altieri and Townsend, 2011). This study represents an approach that associated ERPs with a framework that makes testable and statistically motivated predictions. As a corollary, this framework provides a mechanism to account for capacity changes and co-varying changes in ERPs across listening environments (i.e., facilitatory/inhibitory cross-modal connections during language perception; Altieri and Townsend, 2011; Eidels et al., 2011).

To review, independent models assume that auditory and visual information does not interact as recognition unfolds. Independence predicts that processing capacity/integration efficiency should be approximately equal to 1 across S/N ratios. Violations of independence produced by facilitatory cross-modal interactions elicit a level of efficiency that is greater than 1 (violating the upper bound), while inhibitory interactions yield levels markedly less than 1, and can even approximate fixed capacity [i.e., $C(t) = 1/2$] (Townsend and Wenger, 2004; Eidels et al., 2011; see also Townsend and Nozawa, 1995). A unique feature of capacity is that one can show evidence for different levels of work completed and therefore differences in energy expenditure across time and listening conditions. This differs from other frameworks which conceptualize integration efficiency as an invariant construct unique to a given individual (e.g., Grant et al., 1998; Massaro, 2004).

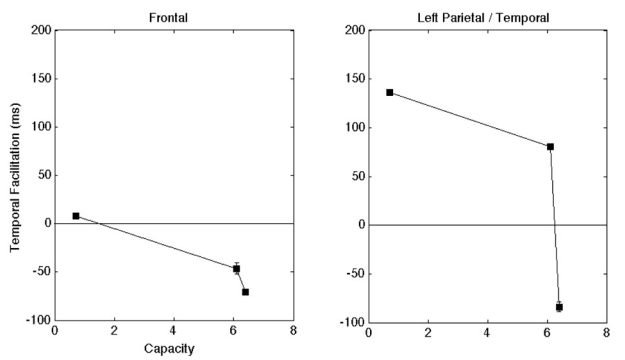


FIGURE 5 | Shows the latency differences in the ERP peak component plotted as a function of capacity for the frontal and left regions. A positive value means that the time-to-peak was faster in the AV compared to the A-only condition. Once again, error bars denote one standard error of the mean computed across individual electrodes (across subjects) in a given region.

Table 3 | Table displaying contrast results for the time to peak for each participant.

Contrast	Late ERP
SUB. 1	
AV − A	−0.42
High	2.90**
Medium	−1.30
Low	−1.16
SUB. 2	
AV − A	−1.85
High	−2.90***
Medium	0.77
Low	3.90***
SUB. 3	
AV − A	0.60
High	0.55
Medium	2.50*
Low	−0.66
SUB. 4	
AV − A	−1.85
High	−5.90***
Medium	−0.77
Low	3.90***

*Negative signs indicate a faster AV relative to A-only time to peak (AV − A), whereas a positive number indicates faster A compared to AV time to peak. Significance (again, ***$p < 0.001$; **$p < 0.01$; and *$p < 0.05$) is shown in each cell corresponding to the high, medium, and low condition.*

The prediction for the late ERP peak component was hypothesized to follow the same qualitative pattern as capacity in the behavioral/RT domain. This should occur due to the low availability of auditory information under degraded listening conditions, which allows complementary visual information to increasingly assist auditory recognition (e.g., Grant et al., 1998; Erber, 2003). Thus, as integration efficiency increases under degraded conditions, the peak amplitude should increase in the audiovisual condition relative to the auditory-only condition (AV > A). We predicted that the neural predictions would covary with the capacity predictions due to the fact that ERP activity may be associated with synchronous neural firing patterns. This is especially true as the A-only amplitude decreases when less phonemic information about manner of articulation is available in the auditory signal. ERP hypotheses were also motivated by findings showing that visual information influences early auditory encoding (e.g., van Wassenhove et al., 2005) and more significantly, the stages of language processing (e.g., Arnal et al., 2011). In a study using magnetoencephalography (MEG), Arnal et al. (2011) (see also Arnal et al., 2009) showed that valid or otherwise congruent audiovisual speech signals were associated with a correlation between a late ERF and an increase in delta frequencies (3–4 Hz). The time-course and MEG scalp topographies indicated that these effects occurred in regions associated with higher language processing. Increasingly useful visual information in lower S/N ratios should lead to more efficient use of visual information in terms of capacity, which ought to be associated with a corresponding increase in a neural index of integration.

Also recall that SR is similar to inverse effectiveness, but differs inasmuch as it assumes that there is an optimal level of noise for achieving maximum multisensory gain. This makes sense in the context of speech perception; if the auditory signal becomes too degraded as discussed previously, then multisensory perception will begin to approximate visual-only performance which is often quite poor (because the auditory signal fails to contain any useful information; Ross et al., 2007; Liu et al., 2013). In a multisensory word recognition task, Liu and colleagues found that the optimal level of AV gain [AV − A]/[1−A] occurred at −12 dB rather than lower S/N ratios. The AV peak amplitude for the time range of 130–200 ms also showed the highest degree of multisensory benefit in the −12 dB condition. One reason we may have observed the highest level of audiovisual gain under the −18 rather than −12 dB condition is that we used a smaller set size of 8 words, which constrained task difficulty.

SUMMARY OF FINDINGS

Integration efficiency was universally inefficient for the high S/N ratio [$C(t) < 1$] but efficient across lower S/N ratios [$C(t) > 1$] (−12 to −18 dB) as predicted.[4] Contrary to intuition, this suggests that multisensory integration may not always be beneficial, particularly for normal-hearing listeners in quiet listening environments. Violations of independent predictions may be observed in the violation of the lower and upper bounds, respectively, in **Figure 2B**. This relation held for each of the four participants. The corresponding audiovisual gain scores [AV−A]/[1−A] and RT gain scores ($A_{RT} - AV_{RT}$) also demonstrated a consistent pattern across each participant, increasing as auditory S/N ratio decreased. These results corroborated recent findings demonstrating the utility and methodological advantages of utilizing capacity as a fine-grained, model-theoretic measure of integration efficiency in speech perception (Altieri, 2010; Altieri and Townsend, 2011). The violations of independence between audiovisual processing as evidenced by $C(t)$ should be somewhat unsurprising given the preponderance of evidence for audiovisual interactions in accuracy (e.g., Massaro, 1987, 2004) and mean ERPs (e.g., Besle et al., 2004; van Wassenhove et al., 2005, 2007; Ponton et al., 2009; Winneke and Phillips, 2011).

Results further demonstrated that as visual information became more useful in lower auditory S/N ratios, capacity values co-varied with significant audiovisual enhancement (AV > A) in the late ERP. These data are consistent with the prediction that visual signals enhance auditory processing in both behavior and in neural processing. An alternative interpretation of this finding is that as auditory noise increases, neural processing as reflected by the EEG increases due to processing difficulty. We argue for the former position because the behavioral and neural data are consistent with the predictions of stochastic resonance (SR). In the clear condition, the average A-only ERP peak was robust, but becomes increasingly attenuated as the auditory signal is degraded by noise. This finding is consistent with ERP research showing evidence for decreased auditory amplitude as noise and processing difficulty increases (Attias et al., 1993; Stevenson et al., 2012). However, when visual information complements the auditory signal, the AV ERP shows a gain relative to the low S/N ratio A-only ERP, thereby reflecting auditory processing in the clear condition.

Finally, the time-to-peak analysis (i.e., $AV_{Time} - A_{Time} > 0$) provided additional support for the hypothesis that visual information interacts with the processing of auditory speech, although the results were somewhat less consistent. The presence of visual speech information is known to speed up auditory processing under different listening conditions (see van Wassenhove et al., 2005). Interestingly, the time-to-peak analysis even showed evidence for inhibition for some participants (i.e., $AV_{Time} > A_{Time}$) in the ERPs when the auditory S/N ratio was lowest (−18 dB). The observed audiovisual slow-down in the time-to-peak analysis presents an intriguing finding. On one hand, degraded listening conditions yield better integration, both in terms of accuracy and capacity, and converging neural evidence was also observed in the ERP analysis. On the other hand, degraded listening conditions led to an audiovisual slow-down

[4] Limited capacity could also result from another situation where the auditory signal-to-noise ratio is low enough that auditory-only recognition accuracy approximates floor performance. Of course, the mechanisms contributing to capacity limitations would be different from cases where auditory recognition is near ceiling. We would probably not implicate cross-channel inhibition as a causal mechanism underlying capacity limitations in such cases. Under this scenario, recognition would be functionally visual-only, where the audiovisual speed and accuracy would probably not differ significantly from the visual-only condition. Some studies of audiovisual gain t have found that efficiency reaches a peak around −12 to −18 dB before tapering off at lower auditory S/N ratios (e.g., Ross et al., 2007).

in terms of the ERP time-to-peak in three participants. The reason for this dissociation is currently unclear. This slowing down of processing in the neural domain may be accounted for by the lack of auditory evidence and the reliance on visually based internal predictions (see van Wassenhove et al., 2005). Since the increase in capacity at lower S/N ratios appears to result from the recruitment of additional resources, it is plausible that the increase in time-to-peak could be due to the cost associated with obtaining extra resources. Crucially, the results from the ERP peak and time-to-peak analyses show evidence for a combined increase in AV amplitude for later processing times in the higher capacity condition relative to the lower capacity conditions. A calculation of the ratio of AV peak amplitude to time-to-peak further indicates that there is more amplitude per unit of time in the highest capacity condition (mean = 15.8) relative to the low (mean = 33.1) [$t_{(142)} = 5.9, p < 0.00001$]. These effects were marginally significant for the lowest capacity condition (i.e., Clear) vs. the −12 dB condition (mean = 18.7) [$t_{(142)} = 1.65, p = 0.10$].

Our findings constitute a significant development by revealing a close correspondence between integration efficiency, as measured by behavior on one hand, and brain signals on the other. Hence, we now have converging evidence for interactive processing in audiovisual speech perception. In this framework (**Figure 1**), auditory and visual information undergo unisensory processing in primary sensory cortices, although linguistic recognition can be enhanced or inhibited via cross-modal connections. Specifically, our combined capacity and neural analysis indicate that these crucial cross-modal interactions occur in later stages during conscious language perception.

REFERENCES

Agus, T. A., Thorpe, S. J., and Pressnitzer, D. (2010). Rapid formation of auditory memories: insights from noise. *Neuron* 66, 610–618. doi: 10.1016/j.neuron.2010.04.014

Altieri, N. (2010). *Toward a Unified Theory of Audiovisual Integration in Speech Perception*. Bloomington: Indiana University.

Altieri, N., Pisoni, D. B., and Townsend, J. T. (2011). Some behavioral and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. *Seeing Perceiving* 24, 513–539. doi: 10.1163/187847611X595864

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

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

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

Attias, J., Urbach, D., Gold, S., and Shemesh, Z. (1993). Auditory event related potentials in chronic tinnitus patients with noise induced hearing loss. *Hear. Res.* 71, 106–113. doi: 10.1016/0378-5955(93)90026-W

Bent, T., Buckwald, 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. A. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: The MIT Press), 153–176.

Bernstein, L. E., Auer, E. T., and Moore, J. K. (2004). "Audiovisual speech binding: convergence or association?" in *Handbook of Multisensory Processing*, eds G. A. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: MIT Press), 203–223.

Berryhill, M., Kveraga, K., Webb, L., and Hughes, H. C. (2007). Multimodal access to verbal name codes. *Percept. Psychophys.* 69, 628–640. doi: 10.3758/BF03193920

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

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

Cappe, C., Thut, G., Romei, V., and Murray, M. (2010). Auditory-visual multisensory interactions in humans: timing, topography, directionality, and sources. *J. Neurosci.* 30, 12572–12580. doi: 10.1523/JNEUROSCI.1099-10.2010

Colonius, H., and Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Front. Integr. Neurosci.* 4:11. doi: 10.3389/fnint.2010.00011

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

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychol. Bull.* 52, 134–140. doi: 10.1037/h0045156

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: a theoretical perspective (L). *J. Acoust. Soc. Am.* 112, 30–33. doi: 10.1121/1.1482076

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

Grice, G. R., Canham, L., and Gwynne, J. W. (1984). Absence of a redundant-signals effect in a reaction time task with divided attention. *Percept. Psychophys.* 36, 565–570. doi: 10.3758/BF03207517

Henkin, Y., Tetin-Schneider, S., Hildesheimer, M., and Kishon-Rabin, L. (2009). Cortical neural activity underlying speech perception in postlingual adult cochlear implant recipients. *Audiol. Neurotol.* 14, 39–53. doi: 10.1159/000153434

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

Liu, B., Lin, Y., Gao, X., and Dang, J. (2013). Correlation between audio-visual enhancement of speech in different noise environments and SNR. A combined behavioral and electrophysiological study. *Neuroscience* 247, 145–151. doi: 10.1016/j.neuroscience.2013.05.007

Martin, B. A., and Stapells, D. R. (2005). Effects of low-pass noise masking on auditory event-related potentials to speech. *Ear Hear.* 26, 195–213. doi: 10.1097/00003446-200504000-00007

Massaro, D. W. (1987). "Speech perception by ear and eye," in *Hearing by Eye: The Psychology of Lip-Reading,* eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum), 53–83.

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.

Mehta, J., Jerger, S., Jerger, J., and Martin, J. (2009). Electrophysiological correlates of word comprehension: event-related potential (ERP) and independent component analysis (ICA). *Int. J. Audiol.* 48, 1–11. doi: 10.1080/14992020802527258

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

Naue, N., Rach, S., Strüber, D., Huster, R. J., Zaehle, T., Körner, U., et al. (2011). Auditory event-related responses in visual cortex modulates subsequent visual responses in humans. *J. Neurosci.* 31, 7729–7736. doi: 10.1523/JNEUROSCI.1076-11.2011

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

Rance, G., Cone-Wesson, B., Wunderlich, J., and Dowell, R. (2002). Speech perception and cortical event related potentials in children with auditory neuropathy. *Ear Hear.* 23, 239–253. doi: 10.1097/00003446-200206000-00008

Riadh, L., Papo, D., Douiri, A., de Bode, S., Gillon-Dowens, M., and Baudonniere, P. (2004). Modulations of 'late' event-related brain potentials in humans by dynamic audiovisual speech stimuli. *Neurosci. Lett.* 372, 74–79. doi: 10.1016/j.neulet.2004.09.039

Rosenblum, L. D. (2005). "Primacy of multimodal speech perception," in *The Handbook of Speech Perception*, eds D. B. Pisoni and R. E. Remez (Malden, MA: Blackwell Publishing), 51–78.

Ross, L. A., Saint-Amour, D., Leavitt, V., Javitt, D. C., and Foxe J. J. (2007). Do you see what I'm saying? optimal visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

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 Department, 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

Stein, B. E., and Meredith, M. A. (1993). The merging of the senses. Cambridge, MA: MIT.

Stevenson, R. A., Bushmakin, M., Kim, S., Wallace, M. T., Puce, A., and James, T. W. (2012). Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. *Brain Topogr.* 25, 308–326. doi: 10.1007/s10548-012-0220-7

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 12–15. doi: 10.1121/1.1907309

Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *The Psychology of Lip-Reading,* eds B. Dodd and R. Campbell (Hillsdale, NJ: LEA), 3–50.

Townsend, J. T., and Ashby, F. G. (1978). "Methods of modeling capacity in simple processing systems," in *Cognitive theory*, Vol. 3, eds J. Castellan and F. Restle (Hillsdale, NJ: Erlbaum), 200–239.

Townsend, J. T., and Ashby, F. G. (1983). *The Stochastic Modeling of Elementary Psychological Processes.* Cambridge: Cambridge University Press.

Townsend, J. T., and Eidels, A. (2011). Workload capacity spaces: a unified methodology for response time measures of efficiency as workload is varied. *Psychon. Bull. Rev.* 18, 659–681. doi: 10.3758/s13423-011-0106-9

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). The serial-parallel dilemma: a case study in a linkage of theory and method. *Psychon. Bull. Rev.* 11, 391–418. doi: 10.3758/BF03196588

van Wassenhove, V., Grant, K., 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

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

Wenger, M. J., and Gibson, B. S. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 708–719. doi: 10.1037/0096-1523.30.4.708

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

Woodward, S. H., Owens, J., and Thompson, L. W. (1990). Word-to-word variation in ERP component latencies: spoken words. *Brain Lang.* 38, 488–503. doi: 10.1016/0093-934X(90)90133-2

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: 26 February 2013; accepted: 22 August 2013; published online: 10 September 2013.

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

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

APPENDIX

	Boat	Date	Gain	Job	Mouse	Page	Shop	Tile
MEAN RTs								
AV_High	1241	1256	1254	1236	1367	1335	1274	1209
AV_Med	1232	1255	1214	1203	1180	1241	1234	1260
AV_Low	1396	1420	1453	1539	1491	1476	1454	1431
A_High	1280	1286	1186	1295	1275	1262	1212	1229
A_Med	1647	1636	1756	2010	1600	2006	1825	1725
A_Low	1853	1888	1953	2070	1893	1988	1948	1993
V-only	1566	1584	1676	1694	1749	1689	1790	1644
ACCURACY								
AV_Med	1.0	0.97	0.97	0.73	0.97	1.0	0.97	0.93
AV_Low	0.97	0.97	0.97	0.64	0.97	0.97	0.97	0.97
A_Med	0.63	0.57	0.78	0.57	0.87	0.82	0.83	0.65
A_Low	0.32	0.25	0.43	0.25	0.23	0.55	0.33	0.29
V_Only	0.93	0.76	0.91	0.59	0.89	0.90	0.75	0.79

Table displaying the mean accuracy and RTs for the 8 words averaged across participants [RT outliers greater than 3000 ms were removed, as was also done in the *C(t)* analyses]. Overall, mean RTs and accuracy were consistent across stimuli within any given condition. Some noticeable exceptions include the mean RTs for "Job" and "Page" relative to the other stimuli in the "A_Med" (−12 dB) condition. One explanation for this particular finding was that the stimuli was that "Job" was often confused with "Shop," leading to lower mean AV and A accuracy (in the −12 and −18 dB conditions) and longer RTs compared the other stimuli. Nonetheless, the same trend of faster audiovisual vs. auditory or visual-only RTs was observed for all stimuli (including "Job"), in the −12 and −18 dB conditions. These mean RT and accuracy data add converging evidence to the capacity and "Gain" data, which indicate considerable audiovisual benefit in lower S/N ratios, and absence of benefit in optimal listening conditions.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 June 2013
doi: 10.3389/fpsyg.2013.00359

Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy

Shahram Moradi[1]*, Björn Lidestam[2] and Jerker Rönnberg[1]

[1] Linnaeus Centre HEAD, Department of Behavioral Sciences and Learning, Linköping University, Linköping, Sweden
[2] Department of Behavioral Sciences and Learning, Linköping University, Linköping, Sweden

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
Axel Winneke, Jacobs University Bremen, Germany
Kaisa Tiippana, University of Helsinki, Finland

***Correspondence:**
Shahram Moradi, Linnaeus Centre HEAD, Department of Behavioral Sciences and Learning, Linköping University, SE-581 83 Linköping, Sweden
e-mail: shahram.moradi@liu.se

This study investigated the degree to which audiovisual presentation (compared to auditory-only presentation) affected isolation point (IPs, the amount of time required for the correct identification of speech stimuli using a gating paradigm) in silence and noise conditions. The study expanded on the findings of Moradi et al. (under revision), using the same stimuli, but presented in an audiovisual instead of an auditory-only manner. The results showed that noise impeded the identification of consonants and words (i.e., delayed IPs and lowered accuracy), but not the identification of final words in sentences. In comparison with the previous study by Moradi et al., it can be concluded that the provision of visual cues expedited IPs and increased the accuracy of speech stimuli identification in both silence and noise. The implication of the results is discussed in terms of models for speech understanding.

Keywords: audiovisual identification, gating paradigm, consonant, word, final word in sentences, silence, noise

INTRODUCTION

The processing of spoken stimuli is interactive. Feed-forward from an incoming signal interacts with feedback from phonological representations in the mental lexicon for the identification of target signals (for a recent review, see Zion Golumbic et al., 2012). For audiovisual speech stimuli, there is additional processing between the incoming auditory and visual signals (see Besle et al., 2008; Lee and Noppeney, 2011). This forms a unified feed-forward signal that interacts with feedback from phonological representations in the mental lexicon [cf. Rapid Automatic Multimodal Binding of PHOnology [RAMBPHO] in the Ease of Language Understanding (ELU) model, Rönnberg et al., 2008]. The multiple interactive processing of audiovisual stimuli results in rapid and highly accurate identification compared with auditory or visual speech alone (Grant et al., 1998). Especially under degraded listening conditions, listeners tend to focus more on the movements of the speaker's face (Buchan et al., 2008). This partially protects the target signal from interference due to acoustic noise by providing information about when and where to expect an auditory signal (Grant, 2001), even though some phonemes and their features may not be readily extractable by vision.

AUDIOVISUAL IDENTIFICATION OF CONSONANTS

Auditory cues provide information about the manner of articulation and voicing, whereas visual cues provide information about the place of articulation (Walden et al., 1975). Correspondence between auditory and visual articulation of phonemes is not one-to-one. Some consonants look the same during visual articulation, such as /k g ŋ/ or /f v/. For instance, the auditory articulation of /b/ results in a clear perception of /b/ in optimum listening condition, while its visual correlates (or visemes) comprise the visual articulation for bilabial consonants /b p m/. The time at which auditory and visual modalities are accessed differs during the audiovisual identification of consonants (Munhall and Tohkura, 1998). Visual information is often available earlier than auditory information (Smeele, 1994).

The audiovisual identification of consonants occurs faster and is more accurate than unimodal auditory or visual presentation (Fort et al., 2010). This is probably due to the accessibility of complementary features associated with using both auditory and visual modalities. van Wassenhove et al. (2005) found that audiovisual speech was processed more quickly than auditory-alone speech. This rapid process was dependent on the degree of visibility of a speech signal; the process was more rapid for highly visible consonants, such as /pa/, than for less visible consonants, such as /ka/. van Wassenhove et al. (2005) proposed an on-line prediction hypothesis to explain how visual and auditory inputs might be combined during the audiovisual identification of speech stimuli. According to their hypothesis, initial visual input first activates phonological representations, and a prediction regarding the identity of the signal is made. This prediction is consistently updated with increasing visual input, and comparisons are made with auditory input in order to solve the identity of a signal. According to Grant and colleagues (Grant and Walden, 1996; Grant et al., 1998), there is little advantage to audiovisual presentation over unimodal presentation if the auditory and visual modalities provide the same critical features, whereas there is a greater advantage when each modality provides different critical features. The greatest advantage of the audiovisual presentation of consonants occurs when the stimuli are presented under noisy conditions (Grant et al., 1998; Jesse and Janse, 2012). Acoustically confusable phoneme pairs, such as /p/ and /k/, can be disambiguated using visual cues (Massaro and Stork, 1998). To conclude, the audiovisual identification of consonants is generally quicker than auditory-alone or visual-alone. As the phonetic cues from either modality act as predictors

for phonetic cues from another modality, more rapid identification of audiovisual presentation would occur than unimodal presentations.

AUDIOVISUAL IDENTIFICATION OF WORDS

Word identification requires an association between an acoustic signal and the phonological-lexical representation in long-term memory (Rönnberg et al., 2008). In the audiovisual identification of words, information from both modalities is combined over time (Tye-Murray et al., 2007), resulting in faster and more accurate identification compared with auditory or visual stimuli alone (Fort et al., 2010). Tye-Murray et al. (2007) proposed the existence of audiovisual neighborhoods composed of overlaps between auditory and visual neighborhoods. According to this view, fewer words exist in the overlap between auditory and visual neighborhoods, resulting in the faster and more accurate identification of audiovisual words. Moreover, the information needed for the identification of vowels, which are the main constituents of words, is available earlier in visual than auditory signals (approximately 160 ms before the acoustic onset of the vowel; Cathiard et al., 1995). In addition, many words are only distinguishable by the place of articulation of one of their constituents (e.g., pet vs. net; Greenberg, 2005). The advantage of audiovisual word identification is more evident under noisy conditions (Sumby and Pollack, 1954; Kaiser et al., 2003; Sommers et al., 2005). Sumby and Pollack (1954) reported that 5–22 dB SNR more noise was tolerated in audiovisual presentation compared to auditory-alone presentation.

COMPREHENSION OF AUDIOVISUAL SENTENCES

In the audiovisual identification of sentences, listeners can benefit from both contextual information and visual cues, resulting in the faster and more accurate identification of target words, especially under degraded listening conditions. The predictability level of sentences is a key factor (Conway et al., 2010); when the auditory signal is degraded, listeners exhibit better performance with highly predictable (HP) audiovisual sentences than with less predictable (LP) ones (Gordon and Allen, 2009). Grant and Seitz (2000) reported that spoken sentences masked by acoustic white noise were recognizable at a lower signal-to-noise ratio (SNR) when the speaker's face was visible. MacLeod and Summerfield (1987, 1990) showed that the provision of visual cues reduced the perceived background noise level by approximately 7–10 dB.

COGNITIVE DEMANDS OF AUDIOVISUAL SPEECH PERCEPTION

Working memory acts as an interface between the incoming signal and phonological representations in semantic long-term memory (Rönnberg et al., 2008). According to the ELU model (Rönnberg et al., 2008), language understanding under optimum listening conditions for people with normal hearing acuity is mostly implicit and effortless. However, under degraded listening conditions (i.e., speech perception in background noise), the demand on the working memory system (including attention and inference-making skills) is increased to help disambiguate the impoverished acoustic signal and match it with corresponding phonological representations in semantic long-term memory. Support for this model comes from studies which show that language understanding under degraded listening conditions is cognitively taxing (for reviews see Rönnberg et al., 2010; Mattys et al., 2012). A recent neuroimaging study demonstrated increased functional connectivity between the auditory (middle temporal gyrus) and inferior frontal gyrus cortices during the perception of auditory speech stimuli in noise (Zekveld et al., 2012; see also Wild et al., 2012), thus suggesting an auditory–cognitive interaction.

Our previous study (Moradi et al., under revision) was in agreement with the ELU model's prediction. The findings showed that working memory and attentional capacities were positively correlated with the early correct identification of consonants and words in noise, while no correlations were found between the cognitive tests and identification of speech tasks in silence. In the noisy condition, listeners presumably are more dependent on their cognitive resources for keeping in mind, testing, and retesting hypothesis. In sum, a combination of auditory and explicit cognitive resources are required in speech perception, but to a lesser extent in silence than in noise.

Adding visual cues to the auditory signal may reduce the working memory load for the processing of audiovisual speech signals for the aforementioned reasons, and there are data to support this (Mousavi et al., 1995; Quail et al., 2009; Brault et al., 2010; Frtusova et al., 2013). Neuroimaging studies have shown that the superior temporal sulcus plays a critical role in audiovisual speech perception in both optimum and degraded listening conditions (Nath and Beauchamp, 2011; Schepers et al., 2013). For instance, Schepers et al. (2013) investigated how auditory noise impacts audiovisual speech processing at three different noise levels (silence, low, and high). Their results showed that auditory noise impacts on the processing of audiovisual speech stimuli in the lateral temporal lobe, encompassing the superior and middle temporal gyri. Visual cues precede auditory information because of natural coarticulatory anticipation, which results in a reduction in signal uncertainty and in the computational demands on brain areas involved in auditory perception (Besle et al., 2004). Visual cues also increase the speed of neural processing in auditory cortices (van Wassenhove et al., 2005; Winneke and Phillips, 2011). Audiological studies have shown that visual speech reduces the auditory detection threshold for concurrent speech sounds (e.g., Grant and Seitz, 2000). This reduction in the auditory threshold makes audiovisual stimuli much easier to detect, thereby reducing the need for explicit cognitive resources (e.g., working memory or attention). Pichora-Fuller (1996) presented sentences with and without background noise and measured the memory span of young adults. The results showed that subjects had better memory span in the audiovisual than in the auditory modality for sentences presented in noise.

Overall, the research indicates that audiovisual speech perception is faster, more accurate, and less effortful than auditory-alone or visual-alone speech perception. By inference, then, audiovisual speech will tax cognitive resources to a lesser extent than auditory-alone speech.

PRESENT STUDY

This study is an extension of that by Moradi et al. (under revision); the same stimuli are used, but are instead presented audiovisually (as compared to auditory-only), using a different sample of participants. The study aimed to determine whether the added visual information would affect the amount of time required for the correct identification of consonants, words, and the final word of HP and LP sentences in both silence and noise using the gating paradigm (Grosjean, 1980). In the gating paradigm, participants hear and see successively increasing parts of speech stimuli until a target is correctly identified; the amount of time required for the correct identification of speech stimuli is termed the isolation point (IP). For example, the participant hears and sees the first 50 ms of a word, then the first 100 ms, and then the first 150 ms and so on, until he or she correctly identifies the word. The participant is required to speculate what the presented stimulus might be after each gate, and is usually also asked to give a confidence rating based on his or her guess. The IP is defined as the duration from the stimulus onset to the point at which correct identification is achieved and maintained without any change in decision after listening to the remainder of the stimulus (Grosjean, 1996).

PREDICTIONS

We predicted that noise would delay the IPs and lower accuracy for the audiovisual identification of consonants and words, which is in line with the findings of our previous study (Moradi et al., under revision). For the audiovisual identification of final words in sentences, listeners can benefit from both the preceding context and visual cues; therefore, we predicted little or no effect of noise on the IPs and accuracy for final word identification in the audiovisual presentation of HP and LP sentences. We also expected that audiovisual presentation would be associated with faster IPs and better accuracy for all gated tasks, compared with auditory presentation alone [which was tested in Moradi et al. (under revision)]. Our previous study (Moradi et al., under revision) also demonstrated significant relationships between explicit cognitive resources (e.g., working memory and attention) and the IPs of consonants and words presented aurally in noise conditions. Specifically, better working memory and attention capacities were associated with the faster identification of consonants and words in noise. In contrast, in the present study, we predicted that the provision of visual cues would aid the identification of consonants and words in noise, and reduce the need for explicit cognitive resources. Hence, we predicted that there would be no significant correlations between the IPs of audiovisual speech tasks in noise and working memory and attention tasks in the present study.

METHODS

PARTICIPANTS

Twenty-four participants (11 men, 13 women) were recruited from the student population of Linköping University. Their ages ranged from 19 to 32 years ($M = 23.3$ years). The students were monolingual Swedish native speakers. All reported having normal hearing and vision (or corrected-to-normal vision), with no psychological or neurological pathology. The participants received 500 SEK (Swedish Kronor) in return for their participation and provided written consent in accordance with the guidelines of the Swedish Research Council, the Regional Ethics Board in Linköping, and the Swedish practice for research on normal populations. It should be noted here that the group of participants in the present study did not differ in their characteristics (i.e., age, gender, educational level, vision and hearing status) with the group of Moradi et al. (under revision).

MEASURES

GATED SPEECH TASKS

A female native speaker of Swedish, looking directly into the camera, read all of the items at a natural articulation rate in a quiet studio. The hair, face, and top part of the speaker's shoulders were visible. She was instructed to begin each utterance with her mouth closed and to avoid blinking while pronouncing the stimuli. Visual recordings were obtained with a RED ONE digital camera (RED Digital Cinema Camera Company, CA) at a rate of 120 frames per second (each frame = 8.33 ms), in 2048 × 1536 pixels. The video recording was edited into separate clips of target stimuli so that the start and end frames of each clip showed a still face.

The auditory stimuli were recorded with a directional electret condenser stereo microphone at 16 bits, with a sampling rate of 48 kHz. The onset time of each auditory target was located as precisely as possible by inspecting the speech waveform using Sound Studio 4 (Felt Tip Inc., NY). Each segmented section was then edited, verified, and saved as a ".wav" file. The root mean square amplitude was computed for each stimulus waveform, and the stimuli were then rescaled to equalize amplitude levels across the different stimuli. A steady-state white noise, borrowed from Hällgren et al. (2006), was resampled and spectrally matched to the speech signals for use as background noise.

Consonants

Eighteen Swedish consonants were used, structured in vowel-consonant-vowel syllable format (/aba, ada, afa, aga, aja, aha, aka, ala, ama, ana, aŋa, apa, ara, aɻa, asa, aʃa, ata, and ava/). The gate size for consonants was set at 16.67 ms. The gating started after the first vowel, /a/, immediately at the start of the consonant onset. Thus, the first gate included the vowel /a/ plus the initial 16.67 ms of the consonant, the second gate added a further 16.67 ms of the consonant (total of 33.33 ms), and so on. The consonant-gating task took 25–40 min per participant to complete.

Words

The words in this study were in consonant-vowel-consonant format, chosen from a pool of Swedish monosyllabic words. The selected words had average to high frequencies according to the Swedish language corpus PAROLE (2011). In total, 46 words were chosen; these were divided into two lists (A and B), each containing 23 words. Both lists were matched in terms of onset phonemes and frequency of use in the Swedish language according to PAROLE (more specifically, each word had three to six alternative words with the same format and pronunciation of the first two phonemes, e.g., the target word /dop/ had the neighbors /dog, dok, don, dos/). For each participant, we presented one list in the silence condition and the

other in the noise condition. The sequence of words was randomized across participants. A pilot study showed that the gate size used for consonants (16.67 ms) led to the subjective feeling that the word-identification task was monotonous, resulting in fatigue and loss of motivation. Therefore, a doubled gate size of 33.3 ms was used for word identification. The first phoneme (consonant) of each word was presented as a whole, and gating was started at the onset of the second phoneme (vowel). The word-gating task took 35–40 min per participant to complete.

Final words in sentences

This study compromised two types of sentences: HP and LP sentences. Predictability was categorized according to the last target word in each sentence which was always a monosyllabic noun (e.g., "Lisa gick till biblioteket för att låna en *bok*"; [Lisa went to the library to borrow a *book*] for an HP sentence; and "Färgen på hans skjorta var *vit*," [The color of his shirt was *white*] for an LP sentence). The predictability of each target word, which was determined on the basis of the preceding words in the sentence, had been assessed in a previous pilot study (Moradi et al., under revision). There were 44 sentences: 22 in each of the HP and LP conditions. The gating started at the onset of the first phoneme of the target word. Due to the supportive effects of the context on word recognition, and based on the pilot data, we set the gate size at 16.67 ms to optimize resolution time. The sentence-gating task took 25–35 min per participant to complete.

HEARING IN NOISE TEST (HINT)

A Swedish version of the Hearing in Noise Test (HINT) (Hällgren et al., 2006), adapted from Nilsson et al. (1994), was used to measure the hearing-in-noise ability of the participant. The HINT sentences consisted of three to seven words. The participants had to repeat each entire sentence correctly in an adaptive ±2 dB SNR. That is, a correct response was followed by a decrease in SNR by 2 dB, and an incorrect response by an increase in SNR by 2 dB. The dependent measure is the calculated SNR (in our case for 50% correct performance). The HINT took approximately 10 min per participant to complete.

COGNITIVE TESTS

Reading span test

In the reading span test (Baddeley et al., 1985), sentences were presented visually, word-by-word in the middle of a computer screen. After each sentence, the participants were instructed to determine whether the sentence was semantically correct or not. After the presentation of a set of sentences, the participants were instructed to repeat either the first word or the last word of each sentence, in correct serial order. Half of the sentences were semantically incorrect, and the other half were semantically correct (Rönnberg, 1990). In this study, two sets of three sentences were initially presented, then two sets of four sentences, followed by two sets of five sentences (for a total of 24 sentences). The reading span score was the aggregated number of words that were correctly recalled across all sentences in the test (maximum score = 24). The reading span test took approximately 15 min per participant to complete.

Paced auditory serial addition test (PASAT)

The PASAT is a test of executive functioning with a strong component of attention (Tombaugh, 2006). The task requires subjects to attend to auditory input, to respond verbally, and to inhibit the encoding of their responses, while simultaneously attending to the next stimulus in a series. Participants were presented with a random series of audio recordings of digits (1–9) and instructed to add pairs of numbers so that each number was added to the number immediately preceding it. This study used the PASAT 2 and PASAT 3 versions of the test (Rao et al., 1991), in which digits were presented at intervals of 2 or 3 s, respectively. The experimenter presented written instructions on how to complete the task, and each participant performed a practice trial. Participants started with PASAT 3, followed by PASAT 2 (faster rate), with a short break between the two tests. The total number of correct responses (maximum possible = 60) at each pace was recorded. The PASAT took approximately 10 min per participant to complete.

SIGNAL-TO-NOISE RATIO (SNR)

In our previous auditory gating study (Moradi et al., under revision), we adjusted the difference between signal and noise to 0 dB. A pilot study for the previous study revealed that very low SNRs resulted in too many errors and SNRs higher than 0 dB were too easy for identification. As the present study was interested in comparing the audiovisual findings with the auditory findings of our previous study (Moradi et al., under revision), we again set the SNR to 0 dB for all audiovisual stimuli.

PROCEDURE

Stimuli were synchronized within 1 ms accuracy and presented using MATLAB (R2009b) and Psychophysics Toolbox (version 3) on an Apple Macintosh computer (Mac Pro 4.1) running OS X (version 10.6.8) (cf. Lidestam, under revision, for more details). The computer was equipped with a fast solid-state hard drive and a fast interface (SATA-III, 6 Gb/s) and graphic card (ATI Radeon HD, 4870 GHz) to assure adequate speed for video rendering and playback. Visual stimuli were displayed in 600 × 600 pixels on a 22″ CRT monitor (Mitsubishi Diamond Pro 2070SB, 120-Hz refresh rate, 800 × 600-pixel resolution) and viewed from a distance of 55 cm. Audio signals were presented binaurally at approximately 65 dB (the range was 62.5–67 dB) via headphones (Sennheiser HDA200), having been adjusted to a comfortable level following the procedure in Moradi et al. (under revision). A second monitor was used for the setup of the experiment; this displayed the MATLAB script and enabled the experimenter to monitor the participants' progress. A screen was placed between the stimulus presentation monitor and the second monitor, preventing participants from seeing the experimenter's screen and response sheets.

The participants were tested individually in a quiet room. Each participant completed all of the gated tasks (consonants, words, and sentences) in one session (the first session), with short rest periods to prevent fatigue. All participants started with the identification of consonants, followed by words and then sentences. The type of listening condition (silence or noise) was counterbalanced across participants such that half of the participants started

with consonant identification in the silence condition, and then proceeded to consonant identification in the noise condition (and vice versa for the other half of the participants). The order of items within each group of consonants, words, and sentences was randomized between participants. The participants were instructed to attend to the auditory speech and the speaker's face on-screen. The participants received written instructions about how to perform the gated tasks, how many sets there were in silence and noise, respectively, and completed several practice trials prior to the main task session. Participants were told to attempt identification after each presentation, regardless of how unsure they were about their identification of the stimulus, but to avoid random guessing. Participants gave their responses aloud, and the experimenter recorded the responses. When necessary, the participants were asked to clarify their responses. The presentation of gates continued until the target was correctly identified on six consecutive presentations. If the target was not correctly identified, stimulus presentation continued until the entire target was presented, even if six or more consecutive responses were identical. The experimenter then started the next trial. When a target was not identified correctly, even after the whole target had been presented, its total duration plus one gate size was used as the estimated IP (cf. Walley et al., 1995; Metsala, 1997; Hardison, 2005). The rationale for this calculated IP was the fact that it is possible some participants give their correct responses at the last gate of a given signal. Hence, estimating an IP equal to the total duration of that speech signal for both correct (even when late) and wrong responses would not be appropriate[1]. There was no specific feedback at any time during the session, except for general encouragement. Furthermore, there was no time pressure for responding to what was heard. The full battery of gating tasks took 85–110 min per participant to complete.

In the second session, the HINT, the reading span test, and the PASAT were administered. The order of the tests was counterbalanced across the participants. The second session took approximately 40 min per participant to complete.

DESIGN

The overall design for the gated tasks, which includes the comparative data from the Moradi et al. (under revision) study, was a 2 × 2 × 4 split-plot factorial design, with Modality as a between participants variable (audiovisual, auditory), combined with the within participant variables: Listening Condition (silence, noise) and Task (consonants, words, LP sentences, HP sentences). For the analysis of the consonant gating task, the design was 2 × 2×18 split-plot factorial: Modality × Listening Condition × Consonant. For the analysis of the word gating task, the design was 2 × 2 split-plot factorial: Modality × Listening Condition. For the final-word-in-sentence gating task, the design was 2 × 2 × 2 split-plot factorial: Modality × Listening Condition × Sentence Predictability.

RESULTS

GATED AUDIOVISUAL TASKS

Table 1 reports the mean responses of participants for the HINT, PASAT 3, PASAT 2, and the reading span test for both the present study and that of Moradi et al. (under revision). There were no significant differences between the two studies for the PASAT 3, PASAT 2, and the reading span test scores. However, the HINT performance was significantly better in the present study than in Moradi et al. (under revision).

Figure 1 shows the mean IPs for the audiovisual gated tasks in both the silence and noise conditions. A two-way repeated-measures analysis (ANOVA) was conducted to compare the means IP for each of the four gated tasks in silence and noise. The results showed a main effect of listening condition, $F_{(1,\ 23)} = 50.69$, $p < 0.001$, $\eta_p^2 = 0.69$, a main effect of the gated tasks, $F_{(1.78,\ 40.91)} = 2898.88$, $p < 0.001$, $\eta_p^2 = 0.99$, and an interaction between listening condition and gated tasks, $F_{(3,\ 69)} = 17.57$, $p < 0.001$, $\eta_p^2 = 0.43$. Four planned comparisons showed that the mean IPs of consonants in silence occurred earlier than in noise, $t_{(23)} = 6.77$, $p < 0.001$. In addition, the mean IPs of words in silence occurred earlier than in noise, $t_{(23)} = 6.09$, $p < 0.001$. However, the mean IPs of final words in HP sentences in silence did not occur earlier than in noise, $t_{(23)} = 0.74$, $p > 0.05$. The same was true for the mean IPs of final words in LP sentences, $t_{(23)} = 0.76$, $p > 0.05$.

Table 2 shows the mean number of correct responses for each of the gated tasks in the silence and noise presented in the audiovisual and auditory modalities. A 2 (Modality: audiovisual vs. auditory) × 2 (Listening Condition: silence vs. noise) × 4 (Gated Task: consonants, words, final words in HP and LP sentences) mixed ANOVA with repeated measures on the second and third factors was conducted to examine the effect of presentation modality on the accuracy for each of four gated tasks. The results showed a main effect of modality, $F_{(1,\ 43)} = 275.32$, $p < 0.001$, $\eta_p^2 = 0.87$, a main effect of listening condition, $F_{(1,\ 43)} = 286.85$, $p < 0.001$, $\eta_p^2 = 0.87$, a main effect of the gated tasks, $F_{(3,\ 129)} = 38.15$, $p < 0.001$, $\eta_p^2 = 0.47$, an interaction between presentation modality and the gated tasks, $F_{(3,\ 129)} = 31.17$, $p < 0.001$, $\eta_p^2 = 0.42$, an interaction between presentation modality and listening condition, $F_{(1,\ 43)} = 145.83$, $p < 0.001$,

Table 1 | Means, *SD* (in parentheses), and significance levels for the HINT and cognitive tests in the present study and in Moradi et al. (under revision).

Type of task	Mean (*SD*) in the present study	Mean (*SD*) in Moradi et al. (under revision)	*p*
HINT	−4.17 (0.72)	−3.11 (1.22)	0.001
PASAT 3	53.38 (4.85)	51.19 (4.38)	0.122
PASAT 2	41.21 (8.33)	40.05 (6.16)	0.602
Reading span test	22.25 (1.67)	21.62 (1.69)	0.216

[1]Similar to Metsala (1997), we also analyzed our data by only including correct responses. There was a main effect of modality, $F_{(1,\ 43)} = 433.41$, $p < 0.001$, $\eta_p^2 = 0.91$; a main effect of listening condition, $F_{(1,\ 43)} = 55.38$, $p < 0.001$, $\eta_p^2 = 0.56$; a main effect of gated tasks, $F_{(2,\ 76)} = 8395.20$, $p < 0.001$, $\eta_p^2 = 0.99$; an interaction between presentation modality and gated tasks, $F_{(3,\ 129)} = 108.60$, $p < 0.001$, $\eta_p^2 = 0.72$; and an interaction between presentation modality and listening condition, $F_{(1,\ 43)} = 20.69$, $p < 0.001$, $\eta_p^2 = 0.33$. However, the three-way interaction between modality, listening condition, and the gated tasks was not significant in this analysis, $F_{(3,\ 41)} = 1.01$, $p > 0.05$, $\eta_p^2 = 0.02$.

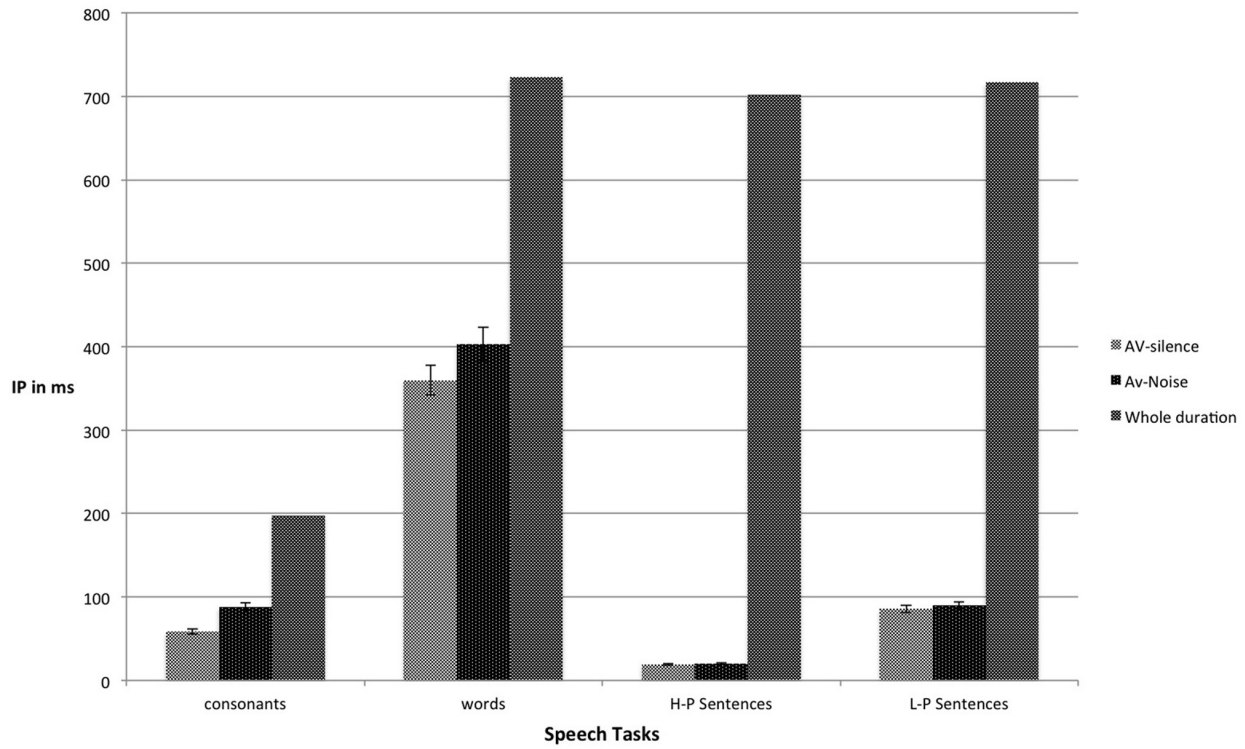


FIGURE 1 | Mean IPs (ms), with accompanying standard errors, for correct identification of audiovisual consonants, words, and final words in HP and LP sentences, in both silence and noise. Whole duration refers to the average total duration from onset to offset.

Table 2 | Accuracy percentages for the identification of gated audiovisual and auditory stimuli: Mean and *SD* (in parentheses).

	Descriptive statistics				Inferential statistics			
	Audiovisual		Auditory		Audiovisual vs. auditory		Silence vs. noise	
Types of gated tasks	Listening condition				Silence (*df* = 43)	Noise (*df* = 43)	Audiovisual (*df* = 23)	Auditory (*df* = 20)
	Silence (a)	Noise (b)	Silence (c)	Noise (d)	(a–c)	(b–d)	(a–b)	(c–d)
Consonants	99.54 (1.58)	89.12 (10.16)	97.35 (3.78)	70.11 (17.52)	$t = 2.59$, $p < 0.013$, $d = 0.76$	$t = 4.52$, $p < 0.001$, $d = 1.33$	$t = 4.85$, $p < 0.001$, $d = 1.37$	$t = 7.50$, $p < 0.001$, $d = 2.21$
Words	100 (0.0)	93.84 (6.77)	96.27 (5.20)	34.58 (17.14)	$t = 3.52$, $p < 0.001$, $d = 1.01$	$t = 15.62$, $p < 0.001$, $d = 4.55$	$t = 4.45$, $p < 0.001$, $d = 0.91$	$t = 15.14$, $p < 0.001$, $d = 4.26$
Final words in LP	100 (0.0)	96.38 (9.90)	87.30 (7.27)	67.06 (20.32)	$t = 8.57$, $p < 0.001$, $d = 2.47$	$t = 6.27$, $p < 0.001$, $d = 1.83$	$t = 1.79$, $p > 0.05$, $d = 0.36$	$t = 4.28$, $p < 0.001$, $d = 1.10$
Final words in HP	99.62 (1.86)	100 (0.0)	94.84 (7.67)	85.71 (7.97)	$t = 2.96$, $p < 0.005$, $d = 0.86$	$t = 8.80$, $p < 0.001$, $d = 2.54$	$t = 1.00$, $p > 0.05$, $d = 0.20$	$t = 2.90$, $p < 0.009$, $d = 1.51$

$\eta_p^2 = 0.77$, and a three-way interaction between modality, listening condition, and the gated tasks, $F_{(3,\ 129)} = 26.27$, $p < 0.001$, $\eta_p^2 = 0.38$. When comparing the accuracy of audiovisual relative to auditory presentation, the greatest advantage of audiovisual presentation was observed for word identification in noise. In the audiovisual modality, noise reduced the accuracy for consonants and words, whereas no effect of noise was found for the accuracy of final words in HP and LP sentences. In the auditory modality, noise reduced the accuracy for all of gated speech tasks. In addition, the most effect of noise on the accuracy in the auditory modality was observed for word identification.

COMPARISON BETWEEN GATED AUDIOVISUAL AND AUDITORY TASKS

The next step in the analysis was to compare the IPs of the audiovisual tasks in the present study with those observed in our previous study (Moradi et al., under revision). This comparison (see **Table 3**) enabled to investigation of the impact that the addition of visual cues had on the amount of time required for the correct identification of stimuli in the auditory gated speech tasks. A 2 (Modality: audiovisual vs. auditory) × 2 (Listening Condition: silence vs. noise) × 4 (Gated Task: consonants, words, final words in HP and LP sentences) mixed ANOVA with repeated measures on the second and third factors was computed to examine the effect of presentation modality on the mean IPs for each of four gated tasks. The results showed a main effect of modality, $F_{(1,\ 43)} = 407.71$, $p < 0.001$, $\eta_p^2 = 0.90$, a main effect of listening condition, $F_{(1,\ 43)} = 282.70$, $p < 0.001$, $\eta_p^2 = 0.87$, a main effect of the gated tasks, $F_{(2,\ 67)} = 2518.60$, $p < 0.001$, $\eta_p^2 = 0.98$, an interaction between presentation modality and the gated tasks, $F_{(3,\ 129)} = 89.21$, $p < 0.001$, $\eta_p^2 = 0.68$, an interaction between presentation modality and listening condition, $F_{(1,\ 43)} = 149.36$, $p < 0.001$, $\eta_p^2 = 0.78$, and a three-way interaction between modality, listening condition, and the gated tasks, $F_{(3,\ 41)} = 40.84$, $p < 0.001$, $\eta_p^2 = 0.49$. When comparing the IPs of audiovisual relative to auditory presentation, the greatest advantage of audiovisual presentation in the silence condition was observed for identification of consonants and words. In the noise condition, the greatest advantage was observed for word identification. Also, when comparing the IPs in the silence condition relative to in the noise condition, the most delaying effect of noise was observed for word identification in the auditory modality. In the audiovisual modality, noise effectively delayed identification of consonants and words, whereas no effect of noise was found for identification of final words in HP and LP sentences.

Consonants

Table 4 shows the mean IPs for the correct identification of consonants in silence and noise presented in the audiovisual and auditory modalities (see also **Figure 2** for the IPs of audiovisual consonants in silence and noise relative to their total durations). A 2 (Modality: audiovisual vs. auditory) × 2 (Listening Condition: silence vs. noise) × 18 (Consonants) mixed ANOVA with repeated measures on the second and third factors was conducted to examine the effect of presentation modality on the IPs for consonant identification. The results showed a main effect of modality, $F_{(1,\ 43)} = 204.50$, $p < 0.001$, $\eta_p^2 = 0.83$, a main effect of listening condition, $F_{(1,\ 41)} = 174.09$, $p < 0.001$, $\eta_p^2 = 0.80$, a main effect for consonants, $F_{(6,\ 273)} = 61.16$, $p < 0.001$, $\eta_p^2 = 0.59$, and a three-way interaction between modality, listening condition, and consonants, $F_{(17,\ 27)} = 2.42$, $p < 0.001$, $\eta_p^2 = 0.05$. Subsequent *t*-test comparisons using a Bonferroni adjustment revealed significant differences ($p < 0.00278$) between silence and noise for /*b f h j k l m n p r ʃ t v*/ within the auditory modality. However, except for /*d k*/, the addition of visual cues did not result in significant differences ($p > 0.00278$) between silence and noise for consonants presented audiovisually. The addition of visual cues did not significantly affect the IPs of /*ŋ ʧ g s*/ in neither silence nor noise, that is, there were no differences between the auditory and audiovisual modalities for these consonants.

Words

A 2 (Modality: audiovisual vs. auditory) × 2 (Listening Condition: silence vs. noise) mixed ANOVA with repeated measures on the second factor was conducted to examine the effect of presentation modality on the IPs for word identification. The results showed a main effect of modality, $F_{(1,\ 43)} = 818.21$, $p < 0.001$, $\eta_p^2 = 0.95$, a main effect of listening condition,

Table 3 | Descriptive and inferential statistics for ips of consonants, words, and final words in HP and LP sentences in silence and noise presented audiovisually and auditorily.

	Descriptive statistics				Inferential statistics			
	Audiovisual		Auditory		Audiovisual vs. auditory		Silence vs. noise	
Types of gated tasks	Listening condition				Silence (*df* = 43)	Noise (*df* = 43)	Audiovisual (*df* = 23)	Auditory (*df* =20)
	Silence (a)	Noise (b)	Silence (c)	Noise (d)	(a–c)	(b–d)	(a–b)	(c–d)
Consonants	58.46 (11.38)	85.01 (19.44)	101.78 (11.47)	161.63 (26.57)	$t = 12.69$, $p < 0.001$, $d = 3.87$	$t = 11.14$, $p < 0.001$, $d = 3.40$	$t = 6.17$, $p < 0.001$, $d = 1.84$	$t = 12.02$, $p < 0.001$, $d = 3.15$
Words	359.78 (25.97)	403.18 (32.06)	461.97 (28.08)	670.51 (37.64)	$t = 12.68$, $p < 0.001$, $d = 3.87$	$t = 25.73$, $p < 0.001$, $d = 7.85$	$t = 6.09$, $p < 0.001$, $d = 1.49$	$t = 17.73$, $p < 0.001$, $d = 6.30$
Final words in LP	85.68 (22.55)	89.94 (15.93)	124.99 (29.09)	305.18 (121.20)	$t = 5.10$, $p < 0.001$, $d = 1.56$	$t = 8.63$, $p < 0.001$, $d = 2.63$	$t = 0.76$, $p > 0.05$, $d = 0.22$	$t = 7.67$, $p < 0.001$, $d = 2.04$
Final words in HP	19.32 (2.69)	19.95 (3.84)	23.96 (3.31)	48.57 (23.01)	$t = 5.18$, $p < 0.001$, $d = 1.58$	$t = 6.01$, $p < 0.001$, $d = 1.83$	$t = 0.74$, $p > 0.05$, $d = 0.19$	$t = 4.96$, $p < 0.001$, $d = 1.50$

Table 4 | Mean IPs, *SD* (in parentheses), and significance levels for the identification of consonants presented audiovisually and auditorily in silence and noise.

Consonants	Modality				*p*			
	Audiovisual		Auditory		Audiovisual vs. auditory		Silence vs. noise	
	Listening condition				Silence	Noise	Audiovisual	Auditory
	Silence (a)	Noise (b)	Silence (c)	Noise (d)	(a–c)	(b–d)	(a–b)	(c–d)
b	50.01 (38.08)	70.15 (44.24)	89.70 (38.19)	157.97 (58.13)	**0.001**	**0.001**	0.069	**0.001**
d	31.96 (23.53)	102.10 (51.86)	138.92 (29.51)	158.76 (25.62)	**0.001**	**0.001**	**0.001**	0.025
f	50.70 (31.28)	59.73 (68.45)	86.53 (17.97)	178.61 (66.92)	**0.001**	**0.001**	0.425	0.001
g	64.60 (37.54)	107.66 (80.92)	146.06 (39.77)	183.37 (47.44)	**0.001**	**0.001**	0.022	0.018
h	75.02 (20.86)	109.05 (57.32)	96.05 (22.31)	186.55 (44.92)	**0.002**	**0.001**	0.007	**0.001**
j	48.62 (22.48)	63.21 (40.23)	66.68 (21.74)	130.18 (41.38)	0.009	**0.001**	0.112	**0.001**
k	27.09 (12.83)	49.32 (25.30)	54.77 (19.11)	85.73 (13.22)	**0.001**	**0.001**	**0.001**	**0.001**
l	46.54 (23.56)	83.35 (72.58)	84.94 (17.41)	176.23 (35.97)	**0.001**	**0.001**	0.014	**0.001**
m	81.96 (31.44)	103.49 (56.69)	79.38 (15.73)	148.44 (72.64)	0.735	0.025	0.044	**0.001**
n	70.15 (48.41)	116.00 (82.62)	105.58 (32.64)	199.25 (61.13)	0.007	**0.001**	0.016	**0.001**
ŋ	100.71 (42.99)	112.52 (72.79)	162.73 (52.16)	169.88 (65.34)	**0.001**	0.008	0.310	0.661
p	22.23 (13.61)	29.17 (26.13)	66.68 (14.91)	111.93 (16.79)	**0.001**	**0.001**	0.226	**0.001**
r	76.40 (25.03)	115.30 (55.38)	88.11 (23.66)	169.88 (34.82)	0.116	**0.001**	0.005	**0.001**
t	136.14 (102.37)	224.35 (156.07)	231.00 (109.60)	338.96 (116.61)	0.004	0.009	0.033	0.008
s	54.18 (11.26)	50.70 (11.51)	68.27 (16.59)	103.99 (65.82)	**0.002**	**0.001**	0.307	0.017
ʃ	45.84 (17.21)	56.26 (43.36)	115.90 (31.84)	166.70 (50.84)	**0.001**	**0.001**	0.295	**0.001**
t	21.53 (10.40)	26.39 (14.68)	44.45 (19.25)	84.94 (13.85)	**0.001**	**0.001**	0.110	**0.001**
v	48.62 (36.10)	51.40 (45.83)	106.37 (47.87)	157.97 (43.67)	**0.001**	**0.001**	0.771	**0.001**

Significant differences according to Bonferroni adjustment ($p < 0.00278$) are in bold.

$F_{(1, 43)} = 354.88$, $p < 0.001$, $\eta_p^2 = 0.89$, and an interaction between modality and listening condition, $F_{(1, 43)} = 152.47$, $p < 0.001$, $\eta_p^2 = 0.78$. One-tailed *t*-tests were subsequently carried out to trace the source of interaction. The results showed that mean audiovisual word identification in silence occurred earlier than mean auditory word identification in silence, $t_{(43)} = 12.68$, $p < 0.001$. In addition, mean audiovisual word identification in noise was earlier than mean auditory word identification in noise, $t_{(43)} = 25.73$, $p < 0.001$. As **Table 3** shows, the difference between silence and noise is larger in the auditory modality than in the audiovisual modality, indicating a less delaying effect of noise in the audiovisual modality.

Final words in sentences

A 2 (Modality: audiovisual vs. auditory) × 2 (Listening Condition: silence vs. noise) × 2 (Sentence Predictability: high vs. low) mixed ANOVA with repeated measures on the second and third factors was conducted to examine the effect of presentation modality on the IPs for final-word identification in sentences. The results showed a main effect of modality, $F_{(1, 43)} = 79.68$, $p < 0.001$, $\eta_p^2 = 0.65$, a main effect of listening condition, $F_{(1, 43)} = 68.11$, $p < 0.001$, $\eta_p^2 = 0.61$, and a main effect of sentence predictability, $F_{(1, 43)} = 347.60$, $p < 0.001$, $\eta_p^2 = 0.89$. There was a three-way interaction between modality, listening condition, and sentence predictability, $F_{(1, 43)} = 53.32$, $p < 0.001$, $\eta_p^2 = 0.55$. Subsequent one-tailed *t*-tests showed that the mean final word identification in both HP and LP sentences occurred earlier in the audiovisual than in the auditory presentation in both silence and noise. As **Table 3** shows, the greatest advantage of audiovisual presentation was observed for final-word identification in LP sentences the in noise condition. In addition, when comparing IPs in silence relative to noise, the most delaying effect of noise was observed for final-word identification in LP sentences in the auditory modality.

CORRELATIONS BETWEEN AUDIOVISUAL GATED TASKS, THE HINT, AND COGNITIVE TESTS

Table 5 shows the Pearson correlations between the IPs for the different gated tasks (lower scores for the gated tasks reflect better performance), the HINT scores (lower scores for the HINT reflect better performance), and the reading span test and PASAT scores (higher scores for the reading span test and PASAT reflect better performance), in both listening conditions (silence and noise). The PASAT 2 was significantly correlated with the HINT and the reading span test. The reading span test was also significantly correlated with the HINT, PASAT 2, and PASAT 3. In addition, the HINT was significantly correlated with IPs of words in noise: the better the participants performed on the HINT, the earlier they could generally identify words presented in noise (and vice versa).

DISCUSSION

IPs FOR THE IDENTIFICATION OF CONSONANTS, WORDS, AND FINAL WORDS IN LP AND HP SENTENCES

Consonants

The mean IPs for consonant identification occurred earlier in silence than in noise (∼58 ms in silence vs. 88 ms in noise),

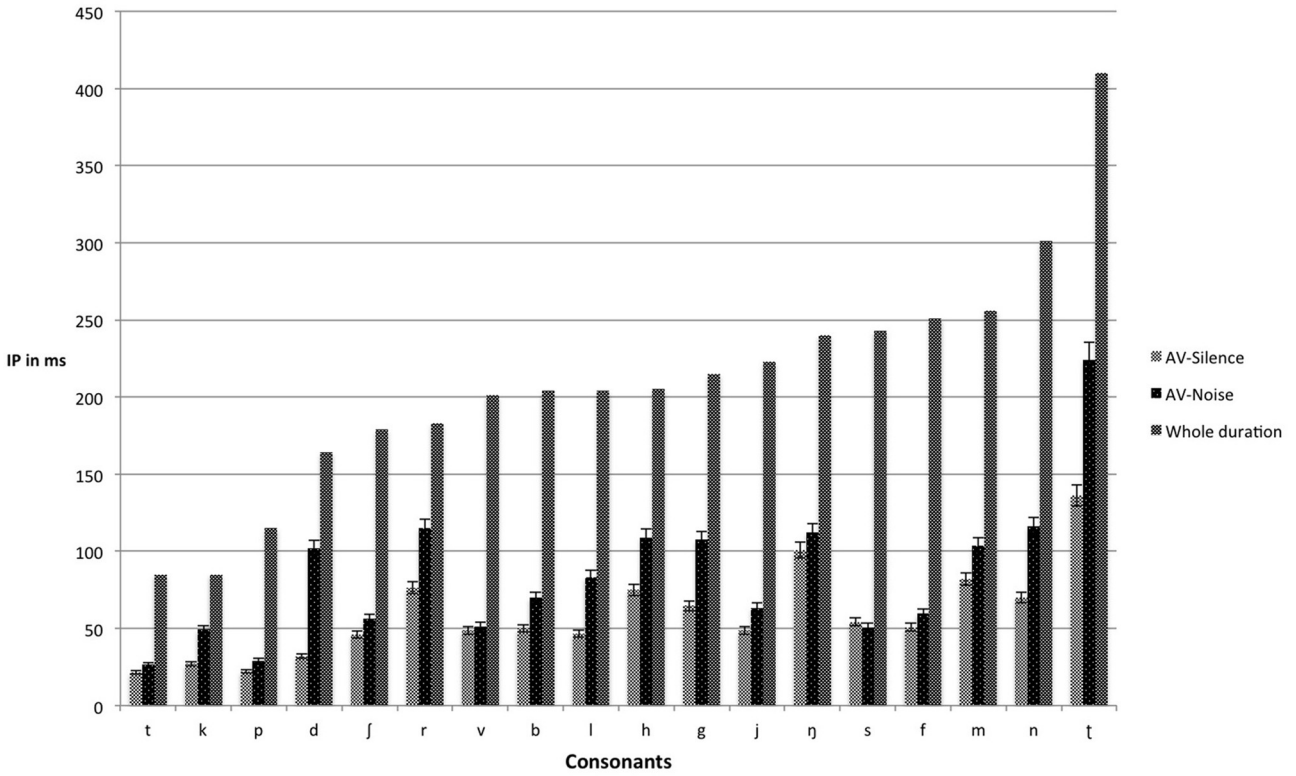


FIGURE 2 | Mean IPs (ms), with accompanying standard errors, for correct identification of audiovisual consonants in both silence and noise. Whole duration refers to the total duration from onset to offset.

Table 5 | Correlations between IPs for the gated audiovisual speech tasks, the HINT, and the cognitive tests.

	1	2	3	4	5	6	7	8	9	10	11	12
1. HINT		−0.34	−0.64**	−0.63**	0.29	0.15	0.12	0.42*	0.15	0.08	−0.04	0.09
2. PASAT 3			0.70**	0.48*	0.09	−0.06	0.01	−0.25	0.06	−0.38	−0.34	−0.07
3. PASAT 2				0.64**	−0.03	0.06	−0.32	−0.27	−0.13	−0.38	−0.15	−0.30
4. RST					−0.05	−0.14	−0.24	−0.40	−0.12	−0.10	0.23	−0.37
5. Consonant-S						0.14	0.09	−0.10	0.29	0.24	0.06	−0.14
6. Consonant-N							−0.34	−0.29	−0.13	−0.18	−0.17	−0.29
7. Word-S								0.29	0.27	0.30	−0.04	0.43
8. Word-N									0.20	0.01	−0.03	0.26
9. HP-S										0.41*	0.21	0.02
10. LP-S											0.54**	0.01
11. HP-N												−0.10
12. LP-N												

Notes: RST, Reading Span Test; Consonant-S, Gated consonant identification in silence; Consonant-N, Gated consonant identification in noise; Word-S, Gated word identification in silence; Word-N, Gated word identification in noise; HP-S, Gated final-word identification in highly predictable sentences in silence; LP-S, Gated final-word identification in less predictable sentences in silence; HP-N, Gated final-word identification in highly predictable sentences in noise; LP-N, Gated final-word identification in less predictable sentences in noise. $*p < 0.05$, $**p < 0.01$.

indicating that noise delayed audiovisual consonant identification. In accordance with the timing hypothesis proposed by van Wassenhove et al. (2005), we hypothesized that background noise would impact on the auditory input of the audiovisual signal, which may make a match between the preceding visual information and the predicted auditory counterparts more difficult, resulting in higher residual errors than in the silence. The resolution of this non-match would require more time (compared with the silence condition) to correctly match the preceding visual signal with the corresponding auditory input. The present

study demonstrated that the amount of time required for the correct identification of consonants was highly variable in both silence and noise (**Figure 2**). The correct identification of consonants was nearly 100% in silence and dropped to 89% in noise (**Table 2**). This is consistent with the findings of Beskow et al. (1997), who reported that listeners correctly identified 76% of Swedish consonants in +3 dB SNR. In sum, our results support our prediction that noise delays IPs and lowers accuracy for the audiovisual identification of consonants.

When comparing the results of consonant identification in the present study with those of Moradi et al. (under revision), it is evident that the provision of visual cues made consonant identification occur earlier in both silence and noise. The results shown in **Table 4** demonstrate that the consonants with the most distinctive visual cues, such as /*b f l m p ʃ t v*/ (cf. Lidestam and Beskow, 2006), were more resistant to noise. However, the added visual cues had no effect on the IPs for /ŋ t *g s*/. Lidestam and Beskow (2006) showed that /t/ was associated with the least visual identification, and /ŋ/ was among the consonants with low identification scores. In terms of accuracy, the correct identification of consonants presented auditory in noise was ∼70% (Moradi et al., under revision). In the current study, this increased to 89% for consonants presented audiovisually. Thus, our findings corroborated the findings of Fort et al. (2010), which showed that audiovisual presentation resulted in higher accuracy and faster identification of phonemes in noise. Our results are also in line with those of Grant and Walden (1996) and Grant et al. (1998) who reported that the visual cues do not need to be very distinctive, as long as they provide cues that are not available from the auditory signal alone, which means that audiovisual identification of consonants in noise is super-additive. In fact, attentional cueing via preceding visual signals provides information about where or when (or where *and* when) the target speech should occur in noisy conditions (Best et al., 2007), which in turn facilitates speech perception in degraded listening conditions. The results were as predicted: audiovisual presentation generally speeded up IPs and improved the accuracy of identified consonants (compared with auditory presentation), and noise generally delayed IPs and lowered accuracy.

Words

The mean IPs for audiovisual word identification in silence occurred earlier than in noise (∼360 ms vs. 403 ms, respectively), which indicates that noise made audiovisual word identification occur later. Audiovisual word identification IPs in noise was correlated with HINT performance (**Table 5**), which indicates that those with a better ability to hear in noise (when not seeing the talker) were also able to identify audiovisual words in noise faster (i.e., when they could see the talker) or vice versa (i.e., those who identified audiovisually presented words in noise early were generally better at hearing in noise when not seeing the talker). **Table 2** shows that the accuracy for correctly identified words in noise was 94%. Our results are in line with those of Ma et al. (2009), who reported the accuracy for word identification to be 90% at 0 dB SNR for monosyllabic English words. Our results are also consistent with the audiovisual gating results of de la Vaux and Massaro (2004), wherein correct word identification at the end of gates was 80% at about +1 dB SNR (they presented stimuli at a maximum of 80% of the total duration of the words). Our results support our prediction that noise delays IPs and reduces accuracy for the audiovisual identification of words.

When comparing the results of the word identification task in the present study with those of our previous study (Moradi et al., under revision), there is an interaction between listening conditions and presentation modality, wherein the impact of noise is reduced in the audiovisual relative to the auditory modality. Audiovisual presentation accelerated word identification to such a degree that the mean IP in audiovisual word identification in noise (403 ms) was less than the mean IP for auditory word identification in silence (462 ms). One explanation as to why auditory word identification takes longer than audiovisual word identification can be inferred from the findings of Jesse and Massaro (2010). They showed that visual speech information is generally fully available early on, whereas auditory speech information is accumulated over time. Hence, early visual speech cues lead to rapid audiovisual word identification. Furthermore, according to Tye-Murray et al. (2007), input received from both auditory and visual channels results in fewer neighborhood candidates (in the overlap of auditory and visual signals) for audiovisual word identification. Together, the results suggest that the time taken to eliminate unrelated candidates when attempting to match an incoming signal with a phonological representation in long-term memory is shorter for words presented audiovisually. This modality protects the speech percept against noise compared to auditory-only presentation. Our results, which showed that the addition of visual cues accelerated lexical access, are consistent with those of Barutchu et al. (2008), Brancazio (2004), and Fort et al. (2010). In our previous study (Moradi et al., under revision), the mean accuracy for word identification in noise was 35%. This increased to 94% in audiovisual word identification in noise in the present study. This result is in line with Potamianos et al. (2001) who reported that at −1.6 dB, the addition of visual cues resulted in 46% improvement in the intelligibility of words presented in noise. As predicted, the results showed that the audiovisual presentation of words resulted in earlier IPs and better accuracy for word identification compared with auditory presentation.

Final words in sentences

As the results show, there was no difference in IPs between silence and noise conditions for final-word identification in HP and LP sentences. The visual cues had a greater compensatory effect for the delay associated with noise than the sentence context had. It did not appear to matter whether the degraded final word was embedded within an HP or LP sentence. The findings are in line with our prediction that noise should not impact significantly on IPs or accuracy for final word identification in HP and LP sentences.

When comparing the results from the present study with those of our previous study (Moradi et al., under revision), The greatest benefit of audiovisual presentation was for LP sentences in noise condition. In sum, there was added benefit associated with the provision of visual cues and the preceding context for the

early decoding of final words in audiovisual sentences in noise. The results were in line with our prediction that audiovisual presentation would result in earlier IPs and better accuracy for final word identification in HP and LP sentences compared with auditory-only presentation.

EFFECT OF MODALITY ON THE HINT PERFORMANCE

It should be noted that there was a significant difference between the HINT performance in the present study and the HINT performance in the study by Moradi et al. (under revision). In both studies, we administered the gated tasks (presented auditory or audiovisually) in the first session and the HINT and cognitive tests in the second session. Audiovisually gated presentation thus seemed to improve HINT performance compared to auditory-only gated presentation. In a study by Bernstein et al. (2013), which examined the impact of audiovisual training on degraded perceptual learning of speech, subjects learned to form paired associations between vocoded spoken nonsense words and nonsense pictures. In one of their experiments, audiovisual training was compared with auditory-only training, and the results showed that, when tested in an auditory-only condition, the audiovisually trained group was better at correctly identifying consonants embedded in nonsense words than the auditory-only group. In other words, auditory-only perception was significantly better following audiovisual training than following auditory-only training. Rosenblum et al. (2007) studied how prior exposure to lip-reading impacts on later auditory speech-in-noise performance. They presented subjects with lip-reading stimuli from the same or a different talker and then measured the auditory speech-in-noise identification performance. The results showed that lip-reading the same talker prior to testing enhanced auditory speech-in-noise performance. Rosenblum et al. hypothesized that the derived amodal idiolectic information from the visual speech of a talker is used to ease auditory speech-in-noise perception. In our studies, the talkers in the gating paradigm and the HINT were not the same but were two different females. To account for this improved HINT performance after audiovisual gating compared to auditory gating, we hypothesize that the cross-modal facilitation, as observed in the HINT scores after audiovisual-gating tasks, can exist even with different talkers to boost the identification of auditory speech-in-noise. According to our findings, we extend the hypothesis by Rosenblum et al. to suggest that visual cues derived from a different talker can still be used to facilitate auditory speech-in-noise function. Further studies are required to see if this cross-modal facilitation from different talkers can be replicated.

COGNITIVE DEMANDS OF AUDIOVISUAL SPEECH PERCEPTION

The current results showed no significant relationships between identification of different audiovisual gated stimuli and performance on cognitive tests, in neither silence nor noise, which supports our prediction that audiovisual speech perception is predominantly effortless. In fact, the audiovisually presentation of speech stimuli reduces working memory load (i.e., Pichora-Fuller, 1996; Frtusova et al., 2013) which in turn eases processing of stimuli especially in noisy condition.

The present study corroborates the findings of our previous study (Moradi et al., under revision) regarding the correlations between the HINT and cognitive tests, such that the HINT was significantly correlated with the reading span test and PASAT 2, suggesting that the subjects with greater hearing-in-noise function had better attention and working memory abilities. When comparing the results from the present study with those of Moradi et al. (under revision), it can be concluded that the identification of audiovisual stimuli (at an equal SNR) demanded less in terms of attention and working memory. This finding is consistent with Fraser et al. (2010), who showed that in the noise condition, speech perception was enhanced and subjectively less effortful for the audiovisual modality than the auditory modality at an equivalent SNR. This is in line with the general prediction made by the ELU model, which states that for relatively poor input signal conditions (i.e., comparing auditory with audiovisual conditions), dependence on working memory and other executive capacities will increase (Rönnberg et al., 2008). We assume that the SNR in the noise condition was not sufficiently demanding to require explicit cognitive resources for the identification of audiovisual speech stimuli in noise; the perceived audiovisual speech signal was well perceived despite the noise. In other words, the audiovisual presentation protected the speech percepts against the noise that has been proven to be an effective masker. It is, however, likely that lower SNRs would increase the demand for explicit cognitive resources.

Our results are not consistent with those of Picou et al. (2011), which showed that low working memory capacity was associated with relatively effortful audiovisual identification of stimuli in noise. It should be noted that Picou et al. (2011) set the SNRs individually for each participant (the audiovisual SNRs ranged from 0 dB to −4 dB, with an average of −2.15 dB across participants). Thus, their method was different to ours, because we used a constant SNR across participants ($SNR = 0$ dB). Hence, the audiovisual task in the noise condition was more difficult in the study of Picou et al. (2011) and probably more cognitively demanding than in our study. Working memory may have been required for the task in the Picou and colleagues' study in order to aid the identification of an impoverished audiovisual signal (cf. the ELU model, Rönnberg et al., 2008). Rudner et al. (2012) showed a significant relationship between working memory capacity and ratings of listening effort for speech perception in noise. Thus, in Picou and colleagues' study, participants with larger working memory capacity may have processed the impoverished audiovisual signal with less effort than those with lower working memory capacity.

One limitation of the present study is that the auditory and audiovisual data stem from different samples, which may raise concerns about potential between-subject sampling errors (although the recruitment and test procedures were identical in both studies). A within-subject design would allow more robust interpretations. Awaiting such an experimental replication, the pattern of results in the current and the previous study by Moradi et al. replicate other independent studies and make theoretical sense. In addition, we used the reading span test and the PASAT with the assumption that they measure amodal working memory and attention capacities of participants. However, there is a

concern about the fact that audiovisual speech tasks and working memory (or attention) was measured separately. In order to draw stronger conclusions about the effect of audiovisual presentation on the working memory (or attention) capacity, a working memory (or attention) task using audiovisual speech stimuli (cf. Frtusova et al., 2013 or Pichora-Fuller, 1996) is proposed for future studies.

CONCLUSIONS

Our results demonstrate that noise significantly delayed the IPs of audiovisually presented consonants and words. However, the IPs of final words in audiovisually presented sentences were not affected by noise, regardless of the sentence predictability level. This suggests that the combination of sentence context and a speech signal with early visual cues resulted in fast and robust lexical activation. In addition, audiovisual presentation seemed to result in fast and robust lexical activation. Importantly, audiovisual presentation resulted in faster and more accurate identification of gated speech stimuli compared to an auditory-only presentation (Moradi et al., under revision).

ACKNOWLEDGMENTS

This research was supported by the Swedish Research Council (349-2007-8654). The authors would like to thank Carl-Fredrik Neikter, Amin Saremi, Mattias Ragnehed, and Niklas Rönnberg for their technical support, and Katarina Marjanovic for speaking the recorded stimuli. The authors would like to thank two anonymous reviewers for their valuable and insightful comments.

REFERENCES

Baddeley, A. D., Logie, R., Nimmo-Smith, I., and Brereton, R. (1985). Components of fluent reading. *J. Mem. Lang.* 24, 119–131. doi: 10.1016/0749-596X(85)90019-1

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., and Crewther, D. P. (2008). When /b/ill with /g/ill becomes /d/ill: evidence for a lexical effect in audiovisual speech perception. *Eur. J. Cogn. Psychol.* 20, 1–11. doi: 10.1080/09541440601125623

Bernstein, L. E., Auer, E. T., 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

Beskow, J., Dahlquist, M., Granström, B., Lundberg, M., Spens, K.-E., and Öhman, T. (1997). "The teleface project: multimodal speech communication for the hearing impaired," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, (Rhodos).

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., and Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recording in human. *J. Neurosci.* 28, 14301–14310. doi: 10.1523/JNEUROSCI.2875-08.2008

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

Best, V., Ozmeral, E. J., and Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *J. Assoc. Res. Otolaryngol.* 8, 294–304. doi: 10.1007/s10162-007-0073-z

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 445–463. doi: 10.1037/0096-1523.30.3.445

Brault, L. M., Gilbert, J. L., Lansing, C. R., McCarley, J. S., and Krmer, A. F. (2010). Bimodal stimulus presentation and expanded auditory bandwidth improve older adults' speech perception. *Hum. Factors* 52, 479–491. doi: 10.1177/0018720810380404

Buchan, J. N., Pare, M., and Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Res.* 1242, 162–171. doi: 10.1016/j.brainres.2008.06.083

Cathiard, M.-A., Lallouache, M. T., Mohamadi, T., and Abry, C. (1995). "Configurational vs. temporal coherence in audio–visual speech perception," in *Proceedings of the 13th International Congress of Phonetic Sciences*, Vol. 3, eds K. Elenius and P. B. Branderud (Stockholm: ICPhS), 218–221.

Conway, C. M., Baurnshmidt, A., Huang, S., and Pisoni, D. B. (2010). Implicit statistical learning in language processing: word predictability is the key. *Cognition* 114, 356–371. doi: 10.1016/j.cognition.2009.10.009

de la Vaux, S. K., and Massaro, D. W. (2004). Audiovisual speech gating: examining information and information processing. *Cogn. Process.* 5, 106–112. doi: 10.1007/s10339-004-0014-2

Fort, M., Spinelli, E., Savariaux, C., and Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech. Commun.* 52, 525–532. doi: 10.1016/j.specom.2010.02.005

Fraser, S., Gange, J. P., Alepins, M., and Dubois, P. (2010). Evaluating the effort expanded to understand speech in noise using a dual-task paradigm: the effects of providing visual cues. *J. Speech. Lang. Hear. Res.* 53, 18–33. doi: 10.1044/1092-4388(2009/08-0140)

Frtusova, J. B., Winneke, A. H., and Phillips, N. A. (2013). ERP evidence that auditory-visual speech facilitates working memory in younger and older adults. *Psychol. Aging.* doi: 10.1037/a0031243. [Epub ahead of print].

Gordon, M. S., and Allen, S. (2009). Audiovisual speech in older and younger adults: integrating a distorted visual signal with speech in noise. *Exp. Aging Res.* 35, 202–219. doi: 10.1080/03610730902720398

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *J. Acoust. Soc. Am.* 109, 2272–2275. doi: 10.1121/1.1362687

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., and Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *J. Acoust. Soc. Am.* 100, 2415–2424. doi: 10.1121/1.417950

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

Greenberg, S. (2005). "A multi–tier framework for understanding spoken language," in *Listening to Speech: An Auditory Perspective*, eds S. Greenberg and W. Ainsworth (Hillsdale, NJ: Erlbaum), 411–433.

Grosjean, F. (1980). Spoken word recognition processes and gating paradigm. *Percept. Psychophys.* 28, 267–283. doi: 10.3758/BF03204386

Grosjean, F. (1996). Gating. *Lang. Cogn. Process.* 11, 597–604. doi: 10.1080/016909696386999

Hällgren, M., Larsby, B., and Arlinger, S. (2006). A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *Int. J. Audiol.* 45, 227–237. doi: 10.1080/14992020500429583

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

Jesse, A., and Janse, E. (2012). Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners. *Lang. Cogn. Process.* 27, 1167–1191. doi: 10.1080/01690965.2011.620335

Jesse, A., and Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Atten. Percept. Psychophys.* 72, 209–225. doi: 10.3758/APP.72.1.209

Kaiser, A. R., Kirk, K. I., Lachs, L., and Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *J. speech. Lang. Hear. Res.* 46, 390–404. doi: 10.1044/1092-4388(2003/032)

Lee, H., and Noppeney, U. (2011). 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

Lidestam, B., and Beskow, J. (2006). Visual phonemic ambiguity and speechreading. *J. Speech. Lang. Hear. Res.* 49, 835–847. doi: 10.1044/1092-4388(2006/059)

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., and Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise. A Bayesian explanation using high-dimensional feature space. *PLoS ONE* 4:e4638. doi: 10.1371/journal.pone.0004638

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

MacLeod, A., and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech perception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br. J. Audiol.* 24, 29–43. doi: 10.3109/03005369009077840

Massaro, D. W., and Stork, D. G. (1998). Speech recognition and sensory integration. *Am. Sci.* 86, 236–244. doi: 10.1511/1998.3.236

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). Speech recognition in adverse listening conditions: a review. *Lang. Cogn. Process.* 27, 953–978. doi: 10.1080/01690965.2012.705006

Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Mem. Cognit.* 25, 47–56. doi: 10.3758/BF03197284

Mousavi, S., Low, R., and Sweller, J. (1995). Reducing cognition load by mixing auditory and visual presentation modes. *J. Educ. Psychol.* 87, 319–334. doi: 10.1037/0022-0663.87.2.319

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

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

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the Hearing In Noise Test (HINT) for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* 95, 338–352. doi: 10.1121/1.408469

Pichora-Fuller, M. K. (1996). "Working memory and speechreadin," in *Speech Reading by Humans and Machines: Models, Systems, and Applications,* eds D. Stork and M. Hennecke (Berlin: Springer-Verlag), 257–274.

Picou, E. M., Kicketts, T. A., and Hornsby, B. W. Y. (2011). Visual cues and listening efforts: individual variability. *J. Speech. Lang. Hear. Res.* 54, 1416–1430. doi: 10.1044/1092-4388(2011/10-0154)

Potamianos, G., Neti, C., Iyengar, G., and Helmuth, E. (2001). "Large-vocabulary audio-visual speech recognition by machines and humans," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, (Aalborg).

Quail, M., Williams, C., and Leitao, S. (2009). Verbal working memory in specific language impairment: the effect of providing visual cues. *Int. J. Speech. Lang. Pathol.* 11, 220–233. doi: 10.1080/17549500802495581

Rao, S. M., Leo, G. J., Bernardin, L., and Unverzagt, F. (1991). Cognitive dysfunction in multiple sclerosis: frequency, patterns, and prediction. *Neurology* 41, 685–691. doi: 10.1212/WNL.41.5.685

Rönnberg, J. (1990). Cognitive and communicative functions: the effects of chronological age and "handicap age". *Eur. J. Cogn. Psychol.* 2, 253–273. doi: 10.1080/09541449008406207

Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *Int. J. Audiol.* 47(Suppl. 2), S171–S177. doi: 10.1080/14992020802301167

Rönnberg, J., Rudner, M., Lunner, T., and Zekveld, A. A. (2010). When cognition kicks in: working memory and speech understanding in noise. *Noise Health* 12, 263–269. doi: 10.4103/1463-1741.70505

Rosenblum, L. D., Miller, R. M., and Sanchez, K. (2007). Lip-read me now, hear me better later. *Psychol. Sci.* 18, 392–396. doi: 10.1111/j.1467-9280.2007.01911.x

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., and Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *J. Am. Acad. Audiol.* 23, 577–589. doi: 10.3766/jaaa.23.7.7

Schepers, I. M., Schneider, T. R., Hipp, J. F., Engel, A. K., Senkowski, D. (2013). Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. *Neuroimage* 70, 101–112. doi: 10.1016/j.neuroimage.2012.11.066

Smeele, P. M. T. (1994). *Perceiving Speech: Integrating Auditory and Visual Speech.* Ph.D. dissertation, Delft: Delft University of Technology.

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

Språkbanken (the Swedish Language Bank). (2011). Available online at: http://spraakbanken.gu.se/

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

Tombaugh, T. N. (2006). A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Arch. Clin. Neuropsychol.* 21, 53–76. doi: 10.1016/j.acn.2005.07.006

Tye-Murray, N., Sommers, M., and Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends. Amplif.* 11, 233–241. doi: 10.1177/1084713807307409

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., Prosek, R. A., and Worthington, D. W. (1975). Auditory and audiovisual feature transmission in hearing-impaired adults. *J. Speech. Lang. Hear. Res.* 18, 272–280.

Walley, A. C., Michela, V. L., and Wood, D. R. (1995). The gating paradigm: effects of presentation format on spoken word recognition by children and adults. *Percept. Psychophys.* 57, 343–351. doi: 10.3758/BF03213059

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

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

Zekveld, A. A., Rudner, M., Johnsrude, I. S., Heslenfeld, D. J., and Rönnberg, J. (2012). Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. *Brain Lang.* 122, 103–113. doi: 10.1016/j.bandl.2012.05.006

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: 28 February 2013; accepted: 31 May 2013; published online: 19 June 2013.

Citation: Moradi S, Lidestam B and Rönnberg J (2013) Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. Front. Psychol. ***4****:359. doi: 10.3389/fpsyg.2013.00359*

This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 03 September 2013
doi: 10.3389/fpsyg.2013.00575

Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes

Annalisa Setti[1,2], Kate E. Burke[1,2], RoseAnne Kenny[1,2,3] and Fiona N. Newell[1,2]*

[1] *Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland*
[2] *TRIL Centre, Trinity College Dublin, Dublin, Ireland*
[3] *Department of Medical Gerontology, Trinity College Dublin, Dublin, Ireland*

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
Thomas C. Gunter, Max Plank Institute for Human Cognitive and Brain Sciences, Germany
MItchell Sommers, Washington University in St. Louis, USA

****Correspondence:***
Fiona N. Newell, Institute of Neuroscience, Trinity College Dublin, Room 3.01, Lloyd Building, Dublin 2, Ireland
e-mail: fnewell@tcd.ie

Recent studies suggest that multisensory integration is enhanced in older adults but it is not known whether this enhancement is solely driven by perceptual processes or affected by cognitive processes. Using the "McGurk illusion," in Experiment 1 we found that audio-visual integration of incongruent audio-visual words was higher in older adults than in younger adults, although the recognition of either audio- or visual-only presented words was the same across groups. In Experiment 2 we tested recall of sentences within which an incongruent audio-visual speech word was embedded. The overall semantic meaning of the sentence was compatible with either one of the unisensory components of the target word and/or with the illusory percept. Older participants recalled more illusory audio-visual words in sentences than younger adults, however, there was no differential effect of word compatibility on recall for the two groups. Our findings suggest that the relatively high susceptibility to the audio-visual speech illusion in older participants is due more to perceptual than cognitive processing.

Keywords: McGurk illusion, audio-visual, speech, multisensory perception, semantic, ageing

INTRODUCTION

Although the human sensory systems are continuously stimulated by multiple sources of information, it is remarkable how the brain efficiently combines the relevant inputs into single objects or events whilst maintaining other sources of information as discrete (see e.g., Calvert et al., 2004). It is known that with ageing, the quality of the sensory inputs diminish due to the degradation of the sensory organs (Fozard and Gordon-Salant, 2001; Gordon-Salant, 2005; Schieber, 2006). Recent research, however, suggests that the ageing brain adapts to these changes to maintain robust perception by relying on the combination of sensory inputs (Laurienti et al., 2006; Peiffer et al., 2007), thus taking advantage of redundancy in cross-sensory information in the environment (Ernst and Bülthoff, 2004). As a consequence, perceptual performance in older persons benefits more from combined inputs than perception in younger adults (Laurienti et al., 2006; Peiffer et al., 2007).

Speech perception is a particularly studied domain in older adults due to its importance for communication and the implications of speech comprehension for social interactions (Pichora-Fuller and Souza, 2003). Since the classic study by Sumby and Pollack (1954) it is well-known that congruent information conveyed across the auditory and visual (i.e., lip-reading) senses facilitates speech perception (Grant and Seitz, 1998; Sommers et al., 2005; Ross et al., 2007; Spehar et al., 2008). In fact visual speech alone can activate the auditory cortex (Calvert et al., 1997). In one recent study comparing younger and older adults, Winneke and Phillips (2011) reported both groups presented reduced P1 and N1 amplitudes in response to audio-visual speech stimuli compared to unisensory stimuli, indicating that fewer resources were necessary to process the audio-visual speech; this effect was more marked in older participants. In addition, both groups showed earlier N1 latency in response to audio-visual stimuli than in unisensory stimuli and the latency shift was related to older adults' hearing thresholds possibly indicating the compensatory function of audio-visual speech to auditory deficits. In fact, older adults, as a consequence of their age-related hearing loss need to rely more on visual speech in order to adequately perceive spoken messages, for example, older adults direct attention toward the speaker's mouth more than younger adults in the attempt to extract sufficient information to support spoken language perception (Thompson and Malloy, 2004) even if lip-reading skills are less efficient in older age (Sommers et al., 2005). While robust evidence shows that older adults benefit more than younger adults of multisensory inputs when speech stimuli are paired with congruent non-speech visual information, e.g., hearing the word "blue" and seeing a blue patch of color (Laurienti et al., 2006), enhanced multisensory integration of audio-visual relative to audio only speech seem not to favor older adults (Sommers et al., 2005; Tye-Murray et al., 2008). This may relate to the quality/integrity of the visual signal as older adults have been shown to have difficulty in processing degraded visual signals (Tye-Murray et al., 2011). For example, Gordon and Allen (2009) showed that older adults benefit from an audio-visual speech input when the level of visual noise is low, however, when the level is high they do not show a benefit possibly because of the difficulty in resolving the visual signal, from the point of view of visual acuity and possibly visual cognitive processing. Nonetheless age-related differences in speech perception appear to be mostly confined to unisensory processing, visual and hearing, while the proportion of benefit

obtained in processing auditory speech when a visual signal is added is similar across age groups (e.g., Sommers et al., 2005).

Clearly, both lower level sensory acuity and higher level cognitive processing play a role in older adults audio-visual processing. In fact older adults can capitalize on visual information in speech, when available, but they can also capitalize on the predictability of the semantic content of the message to support comprehension (Pichora-Fuller, 2008; Sheldon et al., 2008). Higher levels of noise can be tolerated when the semantic content of speech is predictable (Pichora-Fuller, 2008). Indeed when older adults cannot rely on the semantic predictability of a sentence, for example, because the sentence does not express a meaningful content, they benefit more from the addition of visual information to auditory speech than younger adults (Maguinness et al., 2011).

The studies mentioned above utilize either auditory only or congruent audio-visual inputs, in the present study we aimed to assess whether audio visual interactions in speech depend on the semantic content of the spoken message utilizing the McGurk effect. McGurk and MacDonald (1976) reported that auditory syllables (e.g., |ba|) when dubbed over an actor visually articulating a different syllable (e.g., "ga") gave rise to the illusory speech percept of "da." In the first experiment we utilize the "McGurk illusion" to assess whether older adults show enhanced multisensory integration compared to younger adults, and in the second experiment we manipulate the semantic context preceding the illusory audio-visual combination to assess whether the reliance of older adults on semantic predictability determines the susceptibility to the illusion. We hypothesized that older adults would be more susceptible to the illusion than younger adults due to higher susceptibility to multisensory interactions related to non-pathological unisensory decline (but see Tye-Murray et al., 2010).

Factors that affect susceptibility to this illusion have been widely studied (Campbell, 2008 for a review). For example, susceptibility seems to be independent of facial identity, audio-visual co-location and synchrony (Green et al., 1991). Furthermore, the McGurk illusion has been used as an experimental paradigm to investigate efficient audio-visual integration in different populations (de Gelder et al., 2003; Rouger et al., 2008; Bargary et al., 2009). Cienkowski and Carney (2002) previously used the McGurk illusion to compare audio-visual integration across normally hearing older and younger adults. The younger adults comprised two groups; one group was presented with the same auditory stimuli as the older adults whereas the other group was presented with degraded (i.e., with added noise) auditory stimuli, which rendered their auditory perception "equivalent" to the older group. Although all three groups were susceptible to the "McGurk illusion" there was no overall difference across groups in the frequency of reported illusory percepts. However, the lack of a group difference with the particular stimulus set used (the same two syllables repeatedly across the experiment) does not preclude that older adults may show enhanced integration when words as used, as it is known that differences in complexity across syllables, words and sentences as speech units give rise to different perceptual and cognitive processing as shown by the lack of substantial correlation between performance with these different stimuli (see e.g., Sommers et al., 2005).

In the following experiments we investigated susceptibility to the McGurk illusion as a measure of efficient cross-sensory speech perception in older and younger adults. For the purpose of our experiments we used words (Dekle et al., 1992) rather than syllables as stimuli (Alsius et al., 2005). The use of words represents, in our opinion, a more ecological context to the study of multisensory integration of incongruent speech in older adults. The word stimuli contained the relevant phoneme and viseme combination (e.g., |bale|; [gale]) designed to elicit the illusory speech percept (i.e., "dale"). Our paradigm differed, therefore, from previous studies on speech perception which typically measured the benefit of congruent visual inputs on auditory speech (Grant and Seitz, 1998; Grant et al., 1998; Sommers et al., 2005; Tye-Murray et al., 2008). By using incongruent AV stimuli, we can investigate the extent to which speech perception is robust in older adults in an unreliable speech contexts, in which what is heard and what is seen are incongruent.

Speech comprehension has been shown to be dependent on the efficiency in which auditory (speech) and visual (viseme) speech-related information is integrated by the brain. The "McGurk" illusion has recently been extensively used as a tool to investigate how inefficient audio-visual integration is related to impaired speech perception in both the neurotypical population (Jiang and Bernstein, 2011) and in individuals with neural deficits (Woynaroski et al., 2013). Finally, illusions such as the "McGurk" can reveal wider deficits in information processing beyond speech processing (Woynaroski et al., 2013) and therefore offer a powerful, and engaging, tool for the researcher to investigating the processes more "higher-level" functions.

In sum, we hypothesized that older adults would be more susceptible to the McGurk illusion than younger adults (Experiment 1). We also investigated whether a higher occurrence of McGurk illusions in older adults may depend on higher level processing, such as expectations based on semantic context, or lower level perceptual processing. To that end, in Experiment 2 we manipulated the semantic context of an audio-visual sentence such that sentence meaning was either compatible with the combined illusory percept, either of the unisensory components, or both the fused and unisensory components of a target word embedded in the sentence (Windmann, 2004). This allowed us to assess whether expectations based on the semantic context of the sentence play a role in the number of illusions perceived (Windmann, 2004; Ali, 2007) or if the illusion was perceived in a bottom-up, mandatory way irrespective of semantic context (Sams et al., 1998). We predicted that if the illusion was dependent on higher level cognitive processes such as semantic expectations, as it has been suggested that older persons are particularly dependent on semantic context for speech (Pichora-Fuller, 2008), then the frequency of the illusion should be modulated by the relationship between the semantic content of the sentence and the target word more so in older than in younger participants.

EXPERIMENT 1

METHOD

Participants

The final sample for this study was constituted by 26 adult volunteers: 13 younger (mean age of 22 years, $SD = 4$) and 13 older

(mean age of 65.5 years, $SD = 4$) adults. There were 5 male participants in both the younger and older adult groups. All older adults were living independently in the community and were recruited through the Technology Research for Independent Living (TRIL) project (www.trilcentre.org; see Romero-Ortuno et al., 2010 for a characterization of the TRIL cohort).

A larger group of older participants took part in the study as part of a multisensory perception battery of assessments ($n = 37$). Due to the nature of the study, comparing younger and older participants on processing of speech words and sentences, the need to match younger and older on years of education arose. Education has a pervasive effect on cognitive performance and cognitive decline (Stern, 2009) and therefore on language processing. Among our participants 11 had primary education only; 9 had only 2–3 years of secondary education (inter-certificate or other certificate); 12 had secondary education and 4 had college level education or beyond, for 1 participant the education was unknown. Participants with primary-only and intermediate-secondary level of education were excluded as all the younger sample of age >18 had secondary education. That lead to a sample of 16 participants but an appropriate match to younger participants in regard of sex and education was found for 13 of them.

All older participants retained in the final sample had a Mini Mental State Exam (MMSE; Folstein et al., 1975) score higher than 26 (mean = 29, $SD = 1$) indicating normal cognitive function. Vision was either normal or corrected-to-normal (logMAR test mean = 0.05, $SD = 0.05$). Hearing abilities, as assessed through a Hughson Westlake test with Kamplex BA 25 Screening Audiometer, was normal for their age range. Specifically, participants' mean hearing loss at frequency of 3000–4000 Hz was 16.5 dB ($SD = 15$) in the left ear and 15 dB ($SD = 14$) in the right ear. All younger participants reported normal hearing and either normal or corrected-to-normal vision.

The experiments reported here were approved by the St. James Hospital Ethics Committee and the School of Psychology Research Ethics Committee, Trinity College Dublin and conformed to the Declaration of Helsinki. Accordingly, all participants provided informed, written consent prior to taking part in the study.

Stimuli and materials

To create all the stimuli used in the experiment we originally recorded 58 videos [in order to extract 33 visual words (3 repetitions) and 33 auditory words (5 repetitions)] of a female speaker pronouncing a single word. The "McGurk" stimuli were 33 audio-visual incongruent combinations were either taken from a previous stimulus set (Bargary et al., 2009) or were created based on previous literature (e.g., Windmann, 2004) and were known to induce the McGurk illusion (see **Table A1**). The stimuli were created from digital, audio-visual recordings which were taken using a JVC high band digital video camera in a quiet room with natural light illumination. Each audio-visual stimulus was edited using Adobe Premiere® and had duration of, on average, 1 s. The sound was played at 75 dB.

The audio-visual words articulated by the actor were first separated into the audio and visual components to create speech word stimuli which were either auditory only (A-clear/V-degraded), visual only (V-only), AV-congruent or AV-incongruent words (Bargary et al., 2009). Two additional combinations were created to use as practice. In the A-only condition the words used as auditory stimuli were presented together with a masked (i.e., pixelated) version of the corresponding viseme which effectively blurred the visual information but did not remove it (pixilation: average of 6 pixels in the horizontal axis—from ear to ear—and 12 in the vertical axis—from chin to end of forehead-). In the V-only condition the viseme was presented with the auditory word which was masked using white noise. Therefore, although sound was present, it was not related to the speech signal in any way. For the "McGurk illusion" condition, 33 audio-visual combinations were created by combining an incongruent visual word and auditory word such that the time of the lip movements was manually synchronized with the onset and offset of the auditory word by the use of Adobe Premiere®.

Design

The experiment was based on a within-subjects design with the main presentation conditions being either unisensory or multisensory: the two unisensory conditions were A-only and V-degraded and two multisensory conditions were AV-incongruent and AV-congruent. Trials in each condition were presented in separate blocks with four testing blocks in total. Block order was counterbalanced across the entire sample of participants, with the exception of the AV-congruent block which was always presented at the end of the experiment to avoid any effects of congruent word meaning on illusory percepts.

Procedure

Participants were seated in front of a desktop computer with their chin comfortably positioned on a chin-rest at 57 cm from the computer screen. They were informed that they would hear and see an actor pronouncing words and that their task was to report the speech word the actor articulated. The reported word responses were directly recorded by the experimenter onto an electronic file.

At the beginning of each trial a fixation cross appeared at the center of the screen for 700 ms followed by the presentation of the speech word stimulus (A, V, or AV—incongruent and congruent conditions). Participants initiated each trial by pressing the spacebar and there was no time limit for responding.

RESULTS

To assess the task difficulty, we first considered the percentage of trials to which a response was provided by each of the participant groups (i.e., whether correct or incorrect or no response was provided), in each condition (see **Table 1**). In the Aclear/V-degraded condition, the percentage of trials responded to by the older and younger adult groups was 95.3 and 97%, respectively. The V-only condition was considerably more difficult: older and younger participants responded to only 59 and 72% of the trials, and the mean number of trials to which a response was not provided was 8.8 ($SD = 9$) and 10.8, ($SD = 11$), respectively. This difference reflects the relative difficulty that older people have in lip-reading relative to younger adults. There was considerable variation across

Table 1 | Percentage of responses provided and correct responses.

	% responses provided older adults	% responses provided younger adults	% correct older adults	% correct younger adults
A-clear/ V-degraded	95.3	97	15	4.5
V-only	59	72	51	48
AV-congruent	100	100	82	80

participants such that some participants attempted to respond to the majority of trials whereas others responded to none or very few trials. In the AV-incongruent condition (McGurk illusion) the percentage of trials to which a response was provided by the older and younger adults, was 92.5 and 98.6%, respectively. In the congruent AV-condition both older and younger participants responded to all trials. The percentage of words correctly reported was then calculated across participant groups for each condition. The mean percentage of correct responses to the V-only condition was 4.5% ($SD = 5.4$) and 15% ($SD = 26.3$) and to the A-clear/V-degraded condition it was 48% ($SD = 7\%$) and 51% ($SD = 8\%$) for the young and older adult groups, respectively. It is worth noting that the relatively low number of correct A-only responses may be due to the fact that the visual image is only blurred not absent therefore increasing the probability of multisensory interactions in this condition (MacDonald et al., 2000). There was no difference in accuracy between the two groups on the percentage of words correctly reported in either the V-only [$t_{(1,\ 24)} = 1.47$, $p = 0.15$] or the A-clear/V-degraded [$t_{(1,\ 24)} = 1.12$, $p = 0.27$] conditions. In the AV-congruent condition the average percentage of correct responses was 82 and 80% for the older and younger adults, respectively, and there was no difference in performance across the groups [$t_{(1,\ 24)} = 0.53$, $p = 0.6$].

The reported speech words to the AV-incongruent condition were classified as either "McGurk-fused," "McGurk-viseme," "correct-auditory," or "other" responses according to the following criteria: responses to the AV-incongruent stimuli were categorized as "McGurk-fused" response when the reported word corresponded to the fused response; "McGurk-viseme" responses occurred were when the participant reported the visual component of the AV word stimulus; "correct auditory" responses occurred when the participant correctly reported the auditory component (i.e., non-illusory percept); and "other" responses occurred when the participant reported a word that did not correspond to any of the other categories. An example of a "McGurk-fused" response is if the auditory word |bale|when paired with the viseme [kale] produces the reported word of "gale." The "other" category included words which were, for example, similar in phonetics to the auditory word but could not be considered as "McGurk-fused" responses as the place of articulation was the same or similar for the auditory-component of the AV stimulus and the reported word, not intermediate between the place of articulation of the visual and the auditory inputs, as expected in a "fused" response. For example, if the AV combination of |bale|and [kale] gave rise to the unexpected response "bane" this was classified as "other." This "other" category also included unrelated words (e.g., |pin|– [tin] was reported as "elf").

Within the AV-incongruent condition, the percentage of reported words categorized under each of these four response types across older and younger participants was: "correct-auditory," 35 and 37%; McGurk-viseme 7 and 6%; McGurk-fused 37 and 27%; and "other" response was 21 and 29%, respectively.

There were no differences across groups in the "correct-auditory" [$t_{(1,\ 24)} = 0.63$, *n.s.*] and "McGurk-viseme" [$t_{(1,\ 24)} = 0.63$, *n.s.*] conditions. In the "McGurk-fused" condition older participants produced significantly more fused responses than younger participants [$t_{(1,\ 24)} = 3.04$, $p < 0.01$]. The number of reported words which were classified as "other" was significantly higher in younger than in older adults [$t_{(1,\ 24)} = 2.8$, $p < 0.01$]. Furthermore, the overall number of "other" responses was greater than previously reported. One potential reason for this discrepancy may be that different regional accents or languages may influence the extent to which the McGurk illusion is experienced (see e.g., Sekiyama and Tohkura, 1991; Colin et al., 2005).

In order to test whether the effect of group held across different stimuli, we conducted a by item repeated measures ANOVA with proportion of fused responses per item in each group as within items factor. This factor was significant [$F_{(1,\ 32)} = 6.66$, $p < 0.05$], with older adults producing on average a higher proportion of fused responses than younger adults.

In order to assess whether younger and older adults might present different patterns in the responses classified as "other," we conducted further analyses on their types. We found that overall the responses were quite diverse (103 different word types were reported in total across younger and older participants). We classified these responses according to whether they were presented at the same place of articulation as the auditory component of the AV-incongruent stimulus (that is, influenced by the auditory input), as the visual input, or a completely unrelated word. The pattern of response was similar across younger and older participants: we found no difference in the proportion of viseme- (14.75 and 8.22%, respectively) or auditory- (34.4 and 36.9%, respectively) influenced responses across younger and older groups ($\chi^2 = 2.09$, $p = 0.14$). The majority of words in this category were unrelated to either the auditory or viseme of the AV stimulus (with 50.8 and 54.8% of these words provided by the young and older adults, respectively). This shows that while the regional accent of the speaker might have influenced the responses more in younger than in older, there is no specific difference in the kind of "other" responses provided across age groups, and therefore not reflective of a decision bias across the groups.

DISCUSSION

These results show that older participants are more susceptible to the McGurk illusion than younger participants with spoken words. In particular, susceptibility to this illusion appeared to stem from multisensory integration rather than a change in unisensory dominance: group differences existed for the McGurk-fused conditions but not for the McGurk-viseme condition. Moreover, we found no difference across groups in their performance to the unisensory (i.e., A-clear/V-degraded and V-only)

conditions. This lack of difference could be due to the fact that the older adults in this study are relatively high performing individuals from a convenient sample of generally healthy older volunteers (www.trilcentre.org). Importantly older adults in this sample are highly educated and the level of education is generally associated with hearing (e.g., Agrawal et al., 2008), and cognition (e.g., Stern, 2009).

EXPERIMENT 2

INTRODUCTION

Evidence of an effect of semantic context on the McGurk illusion has previously been provided in two studies on younger adults (Windmann, 2004; Ali, 2007). However, in an earlier study an effect of context was not found (Sams et al., 1998) although methodological differences might be the cause of this discrepancy, for example, Sams et al. (1998) used only one kind of auditory-viseme combination, while others presented more varied stimuli (Windmann, 2004). Nevertheless, both of these studies contained methodological advantages which Ali (2007) subsequently adopted in her study. Specifically, Ali manipulated the compatibility of sentence meaning with either the fused word or the unisensory components and reported fewer illusions if sentence meaning was incompatible with the fused percept. None of the previous studies on the role of sentence meaning on susceptibility to the McGurk illusion, however, compared conditions in which either the fused percept or either of its unisensory components were compatible/incompatible with sentence meaning or both. These additional conditions are required to understand whether participants are simply relying on semantic context, i.e., by the semantic compatibility therefore vastly responding in agreement with the compatible percept, or a more bottom-up fusion between the sensory inputs irrespective of context meaning is maintained.

We expected that if speech perception in older adults is driven more by top-down processing than younger adults then their responses should be more dependent on sentence meaning than those of younger participants. As in Experiment 1, we also expected that older participants would experience more illusions overall than younger adults.

METHOD

Participants

See Experiment 1.

Stimuli and materials

The stimuli were digital audio-visual recordings of a female actor articulating sentences. We followed the same procedure as described in Experiment 1 to make these recordings. For the purpose of this experiment, we created 10 target words by pairing together each of 10 auditory words (e.g., |bait|) with one of 10 visemes (e.g., [gate]) in order to produce incongruent word pairings which were most likely to induce the McGurk illusion (e.g., "date"). We then embedded these target words into sentences. For each AV-incongruent word combination we formulated six sentences, and in each of which we manipulated sentence meaning in the following manner. The meaning of the sentence was compatible with either (1) the illusory "McGurk" ("McGurk-fused") percept only, (2) the "McGurk-fused" plus A-clear/V-degraded component, (3) the McGurk-fused plus the V-only component. In the remaining three sentence conditions, the meaning was not compatible with the McGurk-fused percept but was also compatible with one of the unisensory components only, i.e., (4) A-clear/V-degraded or (5) V-only. In the final sentence condition (6) meaning was not compatible with any of the components of the word, fused, or unisensory. **Table 2** provides an illustration of these six sentence conditions based on the specific example of the auditory word |bait|and viseme [gate] pairing.

Prior to the main experiment, these sentences and target word combinations were tested by an independent group of 12 young participants who were instructed to rate, using a 7-point Likert scale, the meaningfulness of each sentence. We also included filler sentences in this rating task for variety. The ratings from these independent judges confirmed our manipulations between sentence meaning and meaning of the target word. In order to assess whether in the completion of each of the sentences there was a bias to produce a given word, we also conducted a sentence completion test with another independent group of 8 participants. They were instructed to complete each of the sentences which was missing the final word. We then calculated how frequently the target word was produced as the final word in each sentence across all participants. The results for meaningful ratings and frequency of word associations are provided in **Table 3** and further discussed in relation to the main study.

Design

The design of the experiment was based on a Group (older vs. younger) by McGurk-fused response compatibility (sentence compatible or incompatible with the "McGurk" word) by compatibility with unisensory response (sentence compatible with the visual, or the auditory input or none of the unisensory components) mixed design. The Group factor was between-subjects whereas McGurk and unisensory compatibility factors were within-subjects factors. The dependent variable was the same as that described in Experiment 1, in that responses to the target word were classified as either "McGurk-fused," McGurk-auditory, Mc-Gurk-viseme or "other."

In the main experiment, one sentence was used as practice and all experimental conditions were presented based on manipulations of the target word in this sentence, yielding 6 practice sentences. The other 9 individual sentences were used as test stimuli (with six different versions of each based on each condition), yielding 54 test sentences in total. The presentation order of the sentences was randomized across participants.

Procedure

Participants were informed that they would be presented with sentences and their task was to repeat the sentence as they had understood it. The experimenter then recorded the sentence reported by participants.

RESULTS

The final word of each reported sentence was categorized based on the same criteria as described in Experiment 1. The proportion of responses within each response category, which were

Table 2 | Example of manipulation of sentence meaning for the audio-visual combination of |bait|and [gate] perceived as "date."

Target word	Stimuli	Sentence	Compatibility
McGurk-only	Auditory	The teenage boy was looking forward to his bait	No
	Visual	The teenage boy was looking forward to his gate	No
	McGurk fusion	The teenage boy was looking forward to his date	Yes
McGurk and Auditory	Auditory	The couple arranged to meet on a bait	No
	Visual	The couple arranged to meet on a gate	Yes
	McGurk fusion	The couple arranged to meet on a date	Yes
McGurk and Visual	Auditory	The fisherman organized his bait	Yes
	Visual	The fisherman organized his gate	No
	McGurk fusion	The fisherman organized his date	Yes
None	Auditory	My phone number was bait	No
	Visual	My phone number was gate	No
	McGurk fusion	My phone number was date	No
Visual-only	Auditory	To catch a trout I need the bait	Yes
	Visual	To catch a trout I need the gate	No
	McGurk fusion	To catch a trout I need the date	No
Auditory-only	Auditory	The bull was locked behind the bait	No
	Visual	The bull was locked behind the gate	Yes
	McGurk fusion	The bull was locked behind the date	No

The target word column indicates the conditions (e.g., sentence's meaning compatible with McGurk fusion and/or the unisensory inputs). The stimuli column indicates, respectively, what the input was in the auditory and visual channel and the expected perceived sentence; the column sentence provides an example. The column headed "Compatibility" indicates whether the final word was compatible or incompatible with the sentence context.

dependent on the compatibility of the AV "McGurk" target word with sentence meaning or on each of the unisensory inputs (i.e., compatible with auditory or visual component), was calculated. These results are plotted in **Figure 1**. We then ran a 2 (compatible with the McGurk percept or not) × 3 unisensory compatibility (no unisensory compatibility or compatible with visual or compatible with auditory component) × 2 Group (older or younger group) mixed design ANOVA. A significant effect of group [$F_{(1,\ 24)} = 5.99$, $p < 0.05$] was found with older participants producing more "McGurk-fused" responses than younger participants (mean proportion of "McGurk-fused" responses were 0.56 and 0.46, respectively). There was also a main effect of compatibility with the McGurk fused word [$F_{(1,\ 24)} = 58.79$, $p < 0.001$] with, on average, more McGurk-fused responses produced when the meaning of the sentence was compatible with the McGurk response (0.59) than when it was not (0.44). Finally there was a main effect of unisensory compatibility [$F_{(1,\ 24)} = 46.5$, $p < 0.001$]: on average, more McGurk responses were produced when none of the unisensory inputs were compatible with sentence meaning (0.62) than when sentence meaning was compatible with either a visual (0.54) or auditory component (0.38; Newman–Keuls *post-hoc*, *p*s < 0.01). None of the interactions between the variables were significant.

When sentence meaning was compatible with both the McGurk fused word (i.e., the target word) and one of the unisensory inputs (either visual or auditory), the proportion of McGurk-fused response was always higher than the amount of auditory or visual based word responses [McGurk compared to Auditory responses in the "McGurk and Auditory" condition: $t = 2.26$, $p < 0.05$; McGurk compared to Visual responses in the "McGurk and Visual" condition: $t = 11.77$, $p < 0.001$]. This result confirms that, while participants were influenced by sentence meaning in responding, they were not entirely driven by it. If it were the case that sentence meaning drove the perception of the target word, then when the sentence was compatible with both the McGurk word and with one of its unisensory components, responses should have been roughly equally distributed between the McGurk-fused and the compatible unisensory response. Instead, we found that participants were responding consistently more accordingly to the (compatible) fused response than to the (compatible) unisensory input.

The results of the word association test lead to some limitations on this conclusion as it appears that the McGurk target word is more likely to be spontaneously associated with the sentence than either of the unisensory words. However, it is worth noting that despite this association, participants still responded by producing the unisensory word, even if it was weakly or not at all associated with the sentence, showing the relevance of the (manipulated) semantic context of the sentence, not of a spontaneous word association. In other words, when sentence meaning was compatible with either the visual or auditory inputs only, participants appeared to respond more in agreement with the meaning than with the unprompted word frequently associated with that sentence.

Table 3 | Mean rating of "meaningfulness" for the sentences in Experiment 2 in each condition.

Compatibility	Input	Mean "meaningfulness" ratings		Frequency of word association (number of target words/number of) produced words)	
		M	*SD*	*M*	*SD*
McGurk only	Auditory	3.2	2.0	0	0
	McGurk	6.6	0.8	2	2.3
	Visual	1.9	1.0	0	0
McGurk and Auditory	Auditory	5.3	1.9	0	0
	McGurk	6.1	1.0	0.5	1.13
	Visual	2.8	2.0	0.2	0.6
McGurk and Visual	Auditory	2.2	0.7	1	2.3
	McGurk	5.1	1.7	0.5	1.1
	Visual	5.5	1.6	0	0
None	Auditory	1.7	0.8	0	0
	McGurk	1.6	0.4	0	0
	Visual	1.6	0.7	0	0
Visual-only	Auditory	2.4	1.3	3	2.8
	McGurk	2.1	0.4	0	0
	Visual	7.0	0.0	0	0
Auditory-only	Auditory	6.6	0.8	1.7	1.7
	McGurk	2.3	1.2	0	0
	Visual	1.9	1.3	0	0

Mean frequency that the final word was spontaneously produced by independent judges per condition.

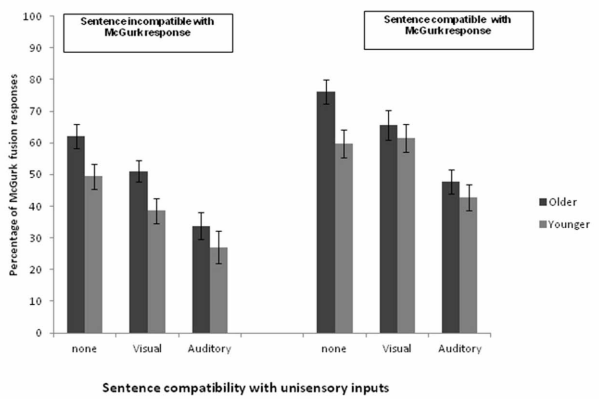


FIGURE 1 | Percentage of "McGurk-fused" responses per condition in Experiment 2. The plots on the left (Sentence incompatible with McGurk response) represent participants' responses in both age groups when sentence meaning was not compatible with the fused word; on the right (compatible with McGurk response) the plots represent the same conditions when the sentence is compatible with the McGurk fusion word.

Considering that some intra- and inter-individual variability is to be expected with McGurk illusion word stimuli, we checked for the one to one correspondence between susceptibility to the illusion in Experiments 1 and 2 (condition where the A and V inputs are compatible with the McGurk fused response only) for the AV pair that we used in both experiments. All older adults showed a 100% by item correspondence, i.e., all items that produced a fused response in Experiment 1 also produced a fusion in Experiment 2 (with the addition of further items producing a fusion in Experiment 2 due to the semantic manipulation as expected). Ten out of thirteen younger adults also showed 100% correspondence and all showed correspondence equal or higher than 60%. A by item analysis between experiments on the average number of illusions in each group revealed a high correlation between experiments both in younger and older participants older $R^2 = 0.6$, $p = 0.02$; younger $R^2 = 0.9$, $p < 0.01$. Although these correlations have to be interpreted with caution due to the limited number of items available for comparison, they suggest a good reliability of the task across experiments.

DISCUSSION

In sum, while older participants were more susceptible than younger adults to the McGurk illusory responses, the effect of sentence meaning on the nature of the target word response (i.e., McGurk-fused, or response based on the auditory or viseme component) did not differ across the two groups.

Our results provided evidence that word perception in both groups was susceptible to the higher-level influence of semantic content of the sentences. However, older adults were more susceptible to the McGurk illusion than younger adults. We did not find a greater influence of context manipulation for older than for younger participants, suggesting that the difference between the age groups on susceptibility to the McGurk illusion was not due to the top-down influence of sentence meaning.

GENERAL DISCUSSION

In the present study we found that older persons are more susceptible to the McGurk audio-visual speech illusion when words and words in sentences are presented than their younger counterparts and that susceptibility to the illusion is influenced but not entirely determined by semantic expectations in relation to meaning.

An age specific benefit of multisensory inputs in older compared with younger adults has been found in the literature when the task requires participants to rely more on one source of information than the other, either because the other has to be ignored, i.e., in selective attention tasks, (e.g., Poliakoff et al., 2006), or because the reliability of one source is higher than the other in some ways (i.e., in incongruent contexts), or else simply because one source provides information which is irrelevant to the task (e.g., background noise Hugenschmidt et al., 2009, see also Mozolic et al., 2010). In line with these considerations the present result shows that when auditory and visual inputs are incongruent, as it is the case in the McGurk illusion, older adults integrate these inputs more often than younger adults. An alternative explanation is that older adults pay more attention to the visual input in order to support their hearing (Thompson and Malloy, 2004), however, this explanation is not fully supported by the

fact that no group difference was found in the visual only condition.

In Experiment 2, we found no interaction between the frequency of McGurk illusions experienced across younger and older groups and the susceptibility to context manipulations. The benefit of semantic compatibility (e.g., more auditory responses provided when the semantic content was congruent with the auditory input) did not differ significantly between younger and older participants. In both groups unisensory semantic biases were associated with a reduction in the number of illusions but not with their complete disappearance. In other words, even when the meaning of the sentence was compatible with either one of the unisensory inputs in the AV incongruent target word, participants were still susceptible to the McGurk illusion.

A limitation of this study is that our final sample of older adults is relatively highly educated, as lower educated elderly were excluded for the purpose of fair comparison with younger adults. Another limitation inherent to the use of audio-visual illusions is the relative inter-individual variability in the susceptibility to the illusion. In addition the relatively high frequency of responses falling in the Other category, due both to the conservative criterion we adopted in classifying the responses provided and the regional accent of the speaker also suggest that these results need to be replicated with a different set of stimuli to ensure their robustness. A further development could also aim to replicate the results with purely unisensory control conditions (e.g., the visual only signal is not accompanied by white noise). Nonetheless this study provides evidence that incongruent audio-visual words are merged more often by older than younger adults (Experiment 1) and this result occurs independently from top-down semantic biases that may favor the fused percept over its unisensory components (Experiment 2). The relationship between this kind of multisensory interaction and congruent language processing needs to the addressed in future studies.

Electrophysiological studies have shown that the McGurk illusion occurs at an early stage of signal processing (Saint-Amour et al., 2007). The left Superior Temporal Sulcus has been shown to play a crucial role in multisensory integration and in susceptibility to the McGurk illusion (Nath and Beauchamp, 2012). However, further studies are necessary to determine the level at which perceptual-semantic interactions occur.

At present, models that allow some contextual constraints on speech perception can account for these results because non-speech information such as visual information and higher level semantic constraints can contribute in recognizing an auditory input (Oden and Massaro, 1978; Massaro and Chen, 2008).

In conclusion the results of the present study suggest that, for the purpose of speech comprehension, older adults combine auditory and visual words more than younger adults, particularly when these words are composed by an incongruent combination of visual and auditory inputs. Importantly, we found in Experiment 2 that while both younger and older participants responded in accordance with semantic compatibility, older adults produced more McGurk illusion responses than younger adults irrespective of the nature of the relationship between sentence meaning and the compatible sensory component of the target word. This result supports the claim that perceptual more than higher level cognitive factors are at the grounds of the higher susceptibility to the McGurk illusion in older relative to younger adults found in the present study.

ACKNOWLEDGMENTS

This research was completed as part of a wider programme of research within the TRIL Centre, (Technology Research for Independent Living). The TRIL Centre is a multi-disciplinary research center, bringing together researchers from UCD, TCD, & Intel, funded by Intel, IDA Ireland and GE Healthcare. www.trilcentre.org. We are grateful to Danuta Lisieska for help with data analyses.

REFERENCES

Agrawal, Y., Platz, E. A., and Niparko, J. K. (2008). Prevalence of hearing loss and differences by demographic characteristics among us adults: data from the national health and nutrition examination survey, 1999-2004. *Arch. Intern. Med.* 168, 1522–1530. doi: 10.1001/archinte.168.14.1522

Ali, A. N. (2007). "Exploring semantic cueing effects using McGurk fusion," in *Auditory-Visual Speech Processing* (Hilvarenbeek: Kasteel Groenendaal). Available online at: http://www.isca-speech.org/archive_open/archive_papers/avsp07/av07_P03.pdf

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

Bargary, G., Barnett, K. J., Mitchell, K. J., and Newell, F. N. (2009). Colored-speech synaesthesia is triggered by multisensory, not unisensory, perception. *Psychol. Sci.* 20, 529–533. doi: 10.1111/j.1467-9280.2009.02338.x

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 593–595. doi: 10.1126/science.276.5312.593

Calvert, G. A., Spence, C., and Stein, B. E. (eds.). (2004). *The Handbook of Multisensory Processes.* Boston, MA: MIT.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 12, 1001–1010. doi: 10.1098/rstb.2007.2155

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

Colin, C., Radeau, 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

de Gelder, B., Vroomen, J., Annen, L., Masthof, E., and Hodiamont, P. (2003). Audio-visual integration in schizophrenia. *Schizophr. Res.* 59, 211–218. doi: 10.1016/S0920-9964(01)00344-9

Dekle, D. J., Fowler, C. A., and Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Percept. Psychophys.* 51, 355–362. doi: 10.3758/BF03211629

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

Folstein, M. F., Folstein, S. E., and McHugh, P. M. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Fozard, J. L., and Gordon-Salant, S. (2001). "Changes in vision and hearing with aging," in *Handbook of the Psychology of Aging*, eds J. E. Birren and K. W. Schaie (San Diego, CA: Academic Press), 241–266.

Gordon, M. S., and Allen, S. (2009). Audio-visual speech in older and younger adults: integrating a distorted visual signal with speech in noise. *Exp. Aging Res.* 35, 202–219. doi: 10.1080/03610730902720398

Gordon-Salant, S. (2005). Hearing loss and aging: new research

Table 3 | Mean rating of "meaningfulness" for the sentences in Experiment 2 in each condition.

Compatibility	Input	Mean "meaningfulness" ratings		Frequency of word association (number of target words/number of) produced words)	
		M	*SD*	*M*	*SD*
McGurk only	Auditory	3.2	2.0	0	0
	McGurk	6.6	0.8	2	2.3
	Visual	1.9	1.0	0	0
McGurk and Auditory	Auditory	5.3	1.9	0	0
	McGurk	6.1	1.0	0.5	1.13
	Visual	2.8	2.0	0.2	0.6
McGurk and Visual	Auditory	2.2	0.7	1	2.3
	McGurk	5.1	1.7	0.5	1.1
	Visual	5.5	1.6	0	0
None	Auditory	1.7	0.8	0	0
	McGurk	1.6	0.4	0	0
	Visual	1.6	0.7	0	0
Visual-only	Auditory	2.4	1.3	3	2.8
	McGurk	2.1	0.4	0	0
	Visual	7.0	0.0	0	0
Auditory-only	Auditory	6.6	0.8	1.7	1.7
	McGurk	2.3	1.2	0	0
	Visual	1.9	1.3	0	0

Mean frequency that the final word was spontaneously produced by independent judges per condition.

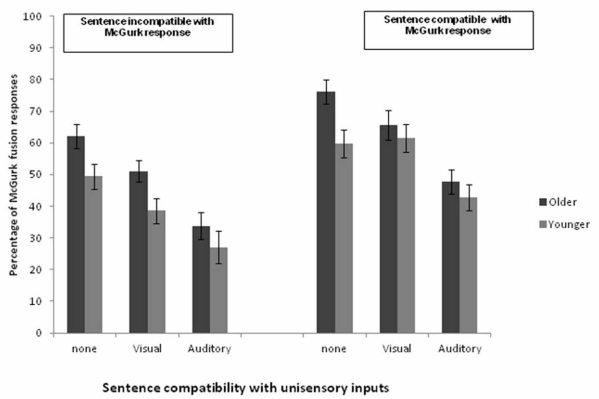


FIGURE 1 | Percentage of "McGurk-fused" responses per condition in Experiment 2. The plots on the left (Sentence incompatible with McGurk response) represent participants' responses in both age groups when sentence meaning was not compatible with the fused word; on the right (compatible with McGurk response) the plots represent the same conditions when the sentence is compatible with the McGurk fusion word.

Considering that some intra- and inter-individual variability is to be expected with McGurk illusion word stimuli, we checked for the one to one correspondence between susceptibility to the illusion in Experiments 1 and 2 (condition where the A and V inputs are compatible with the McGurk fused response only) for the AV pair that we used in both experiments. All older adults showed a 100% by item correspondence, i.e., all items that produced a fused response in Experiment 1 also produced a fusion in Experiment 2 (with the addition of further items producing a fusion in Experiment 2 due to the semantic manipulation as expected). Ten out of thirteen younger adults also showed 100% correspondence and all showed correspondence equal or higher than 60%. A by item analysis between experiments on the average number of illusions in each group revealed a high correlation between experiments both in younger and older participants older $R^2 = 0.6$, $p = 0.02$; younger $R^2 = 0.9$, $p < 0.01$. Although these correlations have to be interpreted with caution due to the limited number of items available for comparison, they suggest a good reliability of the task across experiments.

DISCUSSION

In sum, while older participants were more susceptible than younger adults to the McGurk illusory responses, the effect of sentence meaning on the nature of the target word response (i.e., McGurk-fused, or response based on the auditory or viseme component) did not differ across the two groups.

Our results provided evidence that word perception in both groups was susceptible to the higher-level influence of semantic content of the sentences. However, older adults were more susceptible to the McGurk illusion than younger adults. We did not find a greater influence of context manipulation for older than for younger participants, suggesting that the difference between the age groups on susceptibility to the McGurk illusion was not due to the top-down influence of sentence meaning.

GENERAL DISCUSSION

In the present study we found that older persons are more susceptible to the McGurk audio-visual speech illusion when words and words in sentences are presented than their younger counterparts and that susceptibility to the illusion is influenced but not entirely determined by semantic expectations in relation to meaning.

An age specific benefit of multisensory inputs in older compared with younger adults has been found in the literature when the task requires participants to rely more on one source of information than the other, either because the other has to be ignored, i.e., in selective attention tasks, (e.g., Poliakoff et al., 2006), or because the reliability of one source is higher than the other in some ways (i.e., in incongruent contexts), or else simply because one source provides information which is irrelevant to the task (e.g., background noise Hugenschmidt et al., 2009, see also Mozolic et al., 2010). In line with these considerations the present result shows that when auditory and visual inputs are incongruent, as it is the case in the McGurk illusion, older adults integrate these inputs more often than younger adults. An alternative explanation is that older adults pay more attention to the visual input in order to support their hearing (Thompson and Malloy, 2004), however, this explanation is not fully supported by the

APPENDIX

Table A1 | List of items used in the study.

Audio	Visual	McG
bait	gate	date
bale	cane	gale
bale	kale	gale
been	beep	beam
been	deep	beam
bent	dent	dent
bog	dog	dog
cap	calm	cat
cap	can	cat
clap	can	cat
cop	con	cot
grin	grip	grim
hip	hit	hit
lisp	list	list
mail	day	nail
map	mat	mat
nay	pay	may
neat	peat	meet
pale	tail	kale
pale	trail	kale
pea	tea	key
peek	tea	key
peep	tea	key
pill	tim	kin
pin	tin	kin
pram	cram	cram
ran	rap	ram
rip	rid	rig
shop	shot	shock
shop	shone	shot
veer	dear	gear
vet	get	debt
warn	warp	warm

The acceptable fused responses were either the epected McGurk fusion or a word presenting an intermediate place of articulation between the visual and the auditory word or the same place of articulation as the visual word. Words other than the auditory word that presented the same place of articulation of the auditory word were classified as other.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 16 August 2013
doi: 10.3389/fpsyg.2013.00530

How can audiovisual pathways enhance the temporal resolution of time-compressed speech in blind subjects?

Ingo Hertrich*, Susanne Dietrich and Hermann Ackermann

Department of General Neurology, Center of Neurology, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
Emiliano Ricciardi, University of Pisa, Italy
Nicholas Altieri, Idaho State University, USA

****Correspondence:***
Ingo Hertrich, Department of General Neurology, Center of Neurology, Hertie Institute for Clinical Brain Research, University of Tübingen, Hoppe-Seyler-Strasse 3, D-72076 Tübingen, Germany
e-mail: ingo.hertrich@uni-tuebingen.de

In blind people, the visual channel cannot assist face-to-face communication via lipreading or visual prosody. Nevertheless, the visual system may enhance the evaluation of auditory information due to its cross-links to (1) the auditory system, (2) supramodal representations, and (3) frontal action-related areas. Apart from feedback or top-down support of, for example, the processing of spatial or phonological representations, experimental data have shown that the visual system can impact auditory perception at more basic computational stages such as temporal signal resolution. For example, blind as compared to sighted subjects are more resistant against backward masking, and this ability appears to be associated with activity in visual cortex. Regarding the comprehension of continuous speech, blind subjects can learn to use accelerated text-to-speech systems for "reading" texts at ultra-fast speaking rates (>16 syllables/s), exceeding by far the normal range of 6 syllables/s. A functional magnetic resonance imaging study has shown that this ability, among other brain regions, significantly covaries with BOLD responses in bilateral pulvinar, right visual cortex, and left supplementary motor area. Furthermore, magnetoencephalographic measurements revealed a particular component in right occipital cortex phase-locked to the syllable onsets of accelerated speech. In sighted people, the "bottleneck" for understanding time-compressed speech seems related to higher demands for buffering phonological material and is, presumably, linked to frontal brain structures. On the other hand, the neurophysiological correlates of functions overcoming this bottleneck, seem to depend upon early visual cortex activity. The present Hypothesis and Theory paper outlines a model that aims at binding these data together, based on early cross-modal pathways that are already known from various audiovisual experiments on cross-modal adjustments during space, time, and object recognition.

Keywords: speech perception, blindness, time-compressed speech, audiovisual pathways, speech timing

INTRODUCTION

Speech perception must be considered a multimodal process, arising as an audio-vibrational sensation even prior to birth (Spence and Decasper, 1987) and developing afterward into a primarily audiovisual event. Depending on environmental conditions, lip reading can significantly enhance speech perception (Sumby and Pollack, 1954; Ma et al., 2009). Within this context, the auditory and the visual data streams interact at different – functionally partially independent – computational levels as indicated by various psychophysical effects such as the McGurk and the ventriloquist phenomena (Bishop and Miller, 2011). Furthermore, in combination with cross-modal "equivalence representations" (Meltzoff and Moore, 1997) the visual channel supports early language acquisition, allowing for a direct imitation of mouth movements – based on an innate predisposition for the development of social communication (Streri et al., 2013). Presumably, the underlying mechanism relies on a general action recognition network that is known from primate studies (Buccino et al., 2004; Keysers and Fadiga, 2008), showing that action recognition is closely linked to the motor system, involving a variety of brain structures that have been summarized in a recent review (Molenberghs et al., 2012). In everyday life, the visual channel can be used, first, for the orientation of attention toward the speaking sound source, second, for lipreading, particularly in case of difficult acoustic environments and, third, for visual prosody providing the recipient with additional information related to several aspects of the communication process such as timing, emphasis, valence, or even semantic/pragmatic meaning of spoken language.

Given that speech perception encompasses audiovisual interactions, we must expect significant handicaps at least in early blind subjects with respect to spoken language capabilities. In line with this assumption, delayed speech acquisition has been observed in early blind children (Perez-Pereira and Conti-Ramsden, 1999). By contrast, however, various studies have shown that blind as compared to sighted individuals have superior abilities with respect to auditory perception, compensating at least partially for their visual deficits. Apart from altered central-auditory processing due to intra-modal neural plasticity in both early and late blind subjects (Elbert et al., 2002; Stevens and Weaver, 2009), blind individuals seem, furthermore, to use – at least some components – of their central visual system to support language-related representations

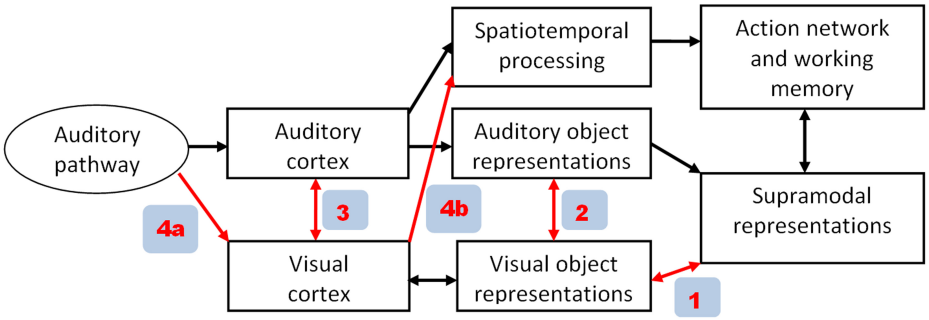


FIGURE 1 | Alternative pathways of visual cortex recruitment during auditory tasks in blind subjects. In the model proposed in the present paper (see **Figure 2**), path 4a/4b plays a major role, enabling visual cortex to process event-timing based on afferent auditory information.

(Röder et al., 2002). In principle, various pathways are available for visual cortex recruitment as shown in **Figure 1**. While particularly in early blind subjects backward projections from supramodal areas (red arrow #1 in **Figure 1**) seem to play a major role for visual cortex activation (Büchel, 2003), more direct pathways among secondary (#2) or primary sensory systems (#3) have also been postulated (Foxe and Schroeder, 2005). In the following we will provide some evidence that even afferent auditory information (#4a) can be utilized by the visual system in blind subjects. This information flow seems to refer to a timing aspect of event recording rather than object recognition (#4b).

Enhanced auditory processing in blind subjects appears to be associated with improved encoding of timing aspects of the acoustic signals. For example, congenitally blind individuals seem to preferentially pay attention to temporal as compared to spatial cues (Röder et al., 2007), and they outperform sighted subjects with respect to temporal resolution capabilities in psychoacoustic backward masking experiments (Stevens and Weaver, 2005). Furthermore, early as well as late blind subjects can acquire the ability to comprehend time-compressed speech at syllable rates up to ca. 20 syllables/s (normal range: ca. 4–8 syllables/s; Moos and Trouvain, 2007). During both backward masking experiments (Stevens et al., 2007) and ultra-fast speech perception (Hertrich et al., 2009, 2013; Dietrich et al., 2013), task performance-related activation of visual cortex has been observed. The aim of this Hypothesis and Theory paper is to delineate potential functional-neuroanatomic mechanisms engaged in enhanced perceptual processing of time-compressed speech in blind subjects. Since this ability has been observed in early as well as late blind individuals (Moos and Trouvain, 2007), we assume that the blind subjects rely on pathways also present in sighted people. However, these connections might not be available for ultra-fast speech processing in the latter group because they are engaged in the processing of actual visual signals.

Against the background of, first, functional magnetic resonance imaging (fMRI) and magnetoencephalographic (MEG) data recorded during the perception of time-compressed speech, second, the literature on cross-modal neuronal pathways in various species and, third, experimental findings dealing with audiovisual illusion effects, a model of visual cortex involvement in ultra-fast speech perception can be inferred. The issue of ultra-fast speech comprehension necessarily touches the question of a more general theory of continuous speech perception in the brain, including all subcomponents such as phonological encoding, lexical access, working memory, and sensorimotor activations of the articulatory system.

NORMAL SPEECH PERCEPTION AND THE TEMPORAL BOTTLENECK

In principle, auditory cortex can follow the temporal envelope of verbal utterances across a wide range of speaking rates (Nourski et al., 2009), indicating that temporal resolution does not represent a limiting factor for the comprehension of time-compressed speech. Thus, we have to assume a "bottleneck" constraining the speed of spoken language encoding. Although the actual execution of motor programs is not required during speech perception, various studies have documented under these conditions the engagement of frontal areas associated with speech production (Pulvermüller et al., 2006). Furthermore, transcranial magnetic stimulation (TMS) experiments revealed these frontal activations to be functionally relevant, e.g., with respect to lexical processing (Kotz et al., 2010; D'Ausilio et al., 2012). Thus, any model of speech perception (e.g., Grimaldi, 2012) has to integrate action-related processing stages bound to the frontal lobe into the cerebral network leading from the acoustic signal to spoken language representations. These cortical areas, subserving, among other things, supramodal operations and transient memory functions, seem to be organized in a more or less parallel manner during speech and music perception (Patel, 2003).

A recent fMRI study (Vagharchakian et al., 2012) suggests that the "bottleneck" in sighted subjects for the comprehension of time-compressed speech arises from limited temporary storage capacities for phonological materials rather than speed constraints of the extraction of acoustic/phonetic features. As a consequence, phonological information might become "overwritten" before it can be fully encoded, a phenomenon contributing, presumably, to backward masking effects. The buffer mechanism for the comprehension of continuous speech has been attributed to left inferior frontal gyrus (IFG), anterior insula, precentral cortex, and upper frontal cortex including the supplementary motor area (SMA and pre-SMA; Vagharchakian et al., 2012). While IFG, anterior insula, and precentral gyrus are supposed to be bound

to mechanisms of speech generation, pre-SMA and SMA might represent an important timing interface between perception- and action-related mechanisms, subserving, among other things, articulatory programming, inner speech, and working memory. More specifically, SMA has been assumed to trigger the execution of motor programs during the control of any motor activities, including speech production. For example, SMA is involved in the temporal organization and sequential performance of complex movement patterns (Tanji, 1994). This mesiofrontal area is closely connected to cortical and subcortical structures that adjust the time of movement initiation to a variety of internal and external demands. In case of acoustically cued simple motor tasks, SMA receives input from auditory cortex, as suggested by a study using Granger causality as a measure of connectivity (Abler et al., 2006). In case of more complex behavior requiring anticipatory synchronization of internal rhythms with external signals such as paced syllable repetitions, SMA seems to also play a major role both in the initiation and the maintenance of motor activity. Furthermore, there seem to be complementary interactions between SMA and the (upper right) cerebellum, the latter being particularly involved in case of increased demands on automation and processing speed during speech production (Riecker et al., 2005; Brendel et al., 2010).

Assuming visual cortex in blind individuals supports temporal signal resolution during speech perception, we have to specify, first, the trigger mechanisms of sighted subjects during perception of normal speech and, second, to delineate how the visual system engages in the encoding of temporal information. Concerning the former issue, Kotz et al. (2009) and Kotz and Schwartze (2010) put forward a comprehensive model of speech perception including an information channel that conveys auditory-prosodic temporal cues via subcortical pathways to pre-SMA and SMA proper. These suggestions also encompass the Asymmetric Sampling in Time hypothesis (Poeppel, 2003; Hickok and Poeppel, 2007) accounting for cortical hemisphere differences that are linked via reciprocal pathways to the cerebellum. As a major focus of the model referred to, Kotz and Schwartze (2010) tried to elucidate the relation of prosodic and syntactic processing – two functional subsystems that have to be coordinated. In analogy to prosody and syntax at the level of the sentence, the syllabic structure of speech, i.e., an aspect of prosody relevant to the timing and relative weighting of segmental phonetic information (Greenberg et al., 2003), provides a temporal grid for the generation of articulation-related speech representations in frontal cortex during perception. In line with the Asymmetric Sampling hypothesis, it has been shown that the syllabic amplitude modulation of the speech envelope is predominantly represented in the right hemisphere (Luo and Poeppel, 2007, 2012; Abrams et al., 2008). Against this background, we hypothesize that a right-hemisphere dominant syllabic timing mechanism is – somehow – linked via SMA to a left-dominant network of phonological processing during speech encoding.

The brain mechanisms combining low-frequency (theta band) syllabic and high-frequency (gamma band) segmental information have been outlined in a recent perspective paper (Giraud and Poeppel, 2012). This model must still be further specified with respect to, first, the pathways connecting right-hemisphere prosodic to left-hemisphere phonetic/phonological representations, second, the involved subcortical mechanisms and, third, the role of SMA for temporal coordination. Considering the salient functional role of syllabicity for speech comprehension (Greenberg et al., 2003), Giraud and Poeppel's model can now be combined with a "syllabic" expansion of the prosodic subcortical-frontal mechanisms including SMA as outlined by Kotz et al. (2009) and Kotz and Schwartze (2010). In this expanded model, a syllable-based representation of speech within the frontal system of spoken language production is temporally coordinated with the incoming speech envelope.

Furthermore, close interactions between frontal speech generation mechanisms and permanent lexical representations have to be postulated since such interactions have also been shown to occur at the level of verbal working memory (Hickok and Poeppel, 2000; Buchsbaum and D'Esposito, 2008; Acheson et al., 2010). Although it must be assumed that verbal working memory, including articulatory loop mechanisms, is based on phonological output structures rather than the respective underlying lexical representations, recent data point at a continuous interaction between articulation-related phonological information and permanent lexical "word node" patterns (Romani et al., 2011). Furthermore, the permanent mental lexicon itself seems to have a dual structure that is linked to the ventral object recognition "what-" pathway within the anterior temporal lobe (phonological features and feature-based word forms; see De Witt and Rauschecker, 2012), on the one hand, and to the dorsal spatiotemporal and more action-related ("where-") projections related to phonological gestures, on the other (Gow, 2012).

Concerning the comprehension of time-compressed speech, syllable rate appears to represent the critical limiting factor rather than missing phonetic information due to shortened segment durations, since insertion of regular silent intervals can largely improve intelligibility in normal subjects (Ghitza and Greenberg, 2009). Since, furthermore, the "bottleneck" seems to be associated with frontal cortex (Vagharchakian et al., 2012), it is tempting to assume that the lack of a syllable-prosodic representation at the level of the SMA limits the processing of time-compressed speech in case syllable rate exceeds a certain threshold. Auditory cortex can, in principle, track the envelope of ultra-fast speaking rates (Nourski et al., 2009) and even monitor considerably higher modulation frequencies, extending into the range of the fundamental frequency of a male speaking voice (Brugge et al., 2009; Hertrich et al., 2012). Furthermore, phase locking to amplitude modulations is consistently stronger within the right than the left hemisphere even at frequencies up to 110 Hz (Hertrich et al., 2004). However, the output from right auditory cortex might have a temporal limitation of syllabic/prosodic event recording: As soon as the modulation frequency approaches the audible range of pitch perception (ca. 16 Hz, that is, for example, the lowest note of an organ) prosodic event recording might compete with a representation of tonal structures. Furthermore, syllable duration at such high speaking rates (16 syllables/s, corresponding to a syllable duration of ca. 60 ms) may interfere with the temporal domain of phonetic features related to voice onset time or formant transitions (ca. 20–70 ms). Thus, the auditory system might not be able to track syllable onsets independently of the extraction of segmental phonological features. Although the segmental (left)

and the prosodic (right) channels could be processed in different hemispheres, the timing of the two auditory cortices might be too tightly coupled in order to separate syllabic from segmental processing if the temporal domains overlap.

A MODEL HOW VISUAL CORTEX IN BLIND SUBJECTS CAN ENHANCE THE PERCEPTION OF TIME-COMPRESSED SPEECH

In this section, a model is presented suggesting right-hemisphere visual cortex activity to contribute to enhanced comprehension of ultra-fast speech in blind subjects. This model is supported, first, by the cortical activation patterns (fMRI, MEG) observed during spoken language understanding after vision loss (see Visual Cortex Involvement in Non-Visual Tasks) and, second, by studies dealing with early mechanisms of signal processing in the afferent audiovisual pathways (see Audiovisual Effects and Associated Pathways). Based, essentially, on the Asymmetric Sampling hypothesis (Poeppel, 2003; Hickok and Poeppel, 2007), the proposed model – as outlined in **Figure 2** – comprises two largely independent data streams, one representing phonological processing including auditory feature recognition in left superior temporal gyrus (STG), frontal speech generation mechanisms, and phonological working memory (green color). The other data stream provides a syllabic timing signal that, in sighted subjects, is predominantly represented at the level of the right-hemisphere auditory system (brown color). The SMA, presumably, synchronizes these two subsystems via subcortical structures (see Kotz and Schwartze, 2010). Blind subjects perceiving ultra-fast speech may use an alternative prosodic channel via an afferent audiovisual pathway including superior colliculus (SC), pulvinar (Pv), and right visual cortex (red arrows). In sighted subjects, these pathways contribute to auditory-driven gating and timing mechanisms for visual object recognition and/or are involved in visual mechanisms of spatial recalibration for auditory events. This afferent signal could provide the visual system with (meaningless) auditory temporal event markers. As a second step, the temporally marked visual events (in sighted) or "empty" visual events (in case of blind subjects) could be transferred to the frontal lobe for further processing such as the timing of inner speech and its encoding into working memory. In sighted subjects, the occipital-frontal pathways, among other things, contribute to the linkage of visually driven motor activity with the temporal structure of visual events.

Synchronization of the left-hemisphere phonological system with the incoming acoustic signal via a prosodic trigger mechanism – that, at an early stage, has some independence from the left-dominant pathway of phonological object recognition – appears to represent an important prerequisite for continuous speech perception under time-critical conditions. This prosodic timing channel, first, might trigger the extraction of phonological features by providing a syllabic grid since the phonological relevance and informational weight of phonological features depends on their position within a syllable (Greenberg et al., 2003). Presumably, transcallosal connections between right and left auditory cortex subserve these functions in sighted people. Second, the syllabic-prosodic timing signal could coordinate frontal speech generation and working memory mechanisms with the auditory input signal since speech generation is organized in a syllabic output structure. In particular, these interactions are important for the exact timing of top-down driven forward predictions with regard to the expected acoustic speech signal. Thus, the presence of a syllabic timing signal can significantly enhance the utilization of informational redundancy (predictability) during continuous realtime speech perception. It should also be mentioned that, although we assume an early signal-driven mechanism, visual cortex activation was found to be considerably weaker in case of (unintelligible) backward as compared to forward speech (Dietrich et al., 2013; Hertrich et al., 2013). We have to assume, thus, that top-down mechanisms providing information on the meaningfulness of the sound signal – arising, presumably, within frontal cortex – have an impact on the recruitment of the visual cortex during ultra-fast speech comprehension. Particularly, such interactions might be relevant for functional neuroplasticity processes during the training phase when blind subjects learn to accelerate their speech perception system using visual resources.

Apart from right-hemisphere mechanisms of prosody encoding, blind subjects seem also to engage ventral aspects (fusiform gyrus, FG) of their left-hemisphere visual system during ultra-fast speech perception (Hertrich et al., 2009; Dietrich et al., 2013). Therefore, left FG was added to **Figure 2** although the functional role of this occipito-temporal area remains to be further specified. At least parts of left FG appear to serve as a secondary phonological and/or visual word form area, linked to the left-hemisphere language processing network (McCandliss et al., 2003; Cao et al., 2008; Cone et al., 2008; Dietrich et al., 2013).

VISUAL CORTEX INVOLVEMENT IN NON-VISUAL TASKS

A large number of studies report visual cortex activity in blind subjects during non-visual tasks, but the functional relevance of these observations is still a matter of debate (Röder et al., 2002; Burton, 2003; Burton et al., 2010; Kupers et al., 2011). Most studies (see Noppeney, 2007 for a comprehensive review) focus on early blind subjects, reporting visual cortex activity related to various tasks such as linguistic processing or braille reading. In some cases, a causal relationship has explicitly been demonstrated, e.g., by means of TMS showing that a transient "virtual lesion" in left occipital cortex interferes with semantic verbal processing (Amedi et al., 2004).

Regarding the neuronal mechanisms of functional cross-modal plasticity, cortico-cortical connections have been hypothesized on the basis of animal experiments, either direct cross-modal connections between, e.g., auditory and visual cortex, or backward projections from higher-order supramodal centers toward secondary and primary sensory areas (see e.g., Foxe and Schroeder, 2005; Bavelier and Hirshorn, 2010). Thereby, even in congenitally blind subjects, the supramodal representations seem to be quite similarly organized as in sighted individuals, indicating that supramodal representations form a stable pattern, largely independent of input modality (Ricciardi and Pietrini, 2011). In most examples of the engagement of the central visual system in blind subjects during non-visual cognitive tasks such as linguistic processing, thus, a top-down mode of stimulus processing from higher-order representations toward visual cortex has been assumed (Büchel et al., 1998; Büchel, 2003; Macaluso and Driver, 2005). By contrast, functional neuroplasticity via subcortical pathways has rarely been taken into account (Bavelier and Neville,

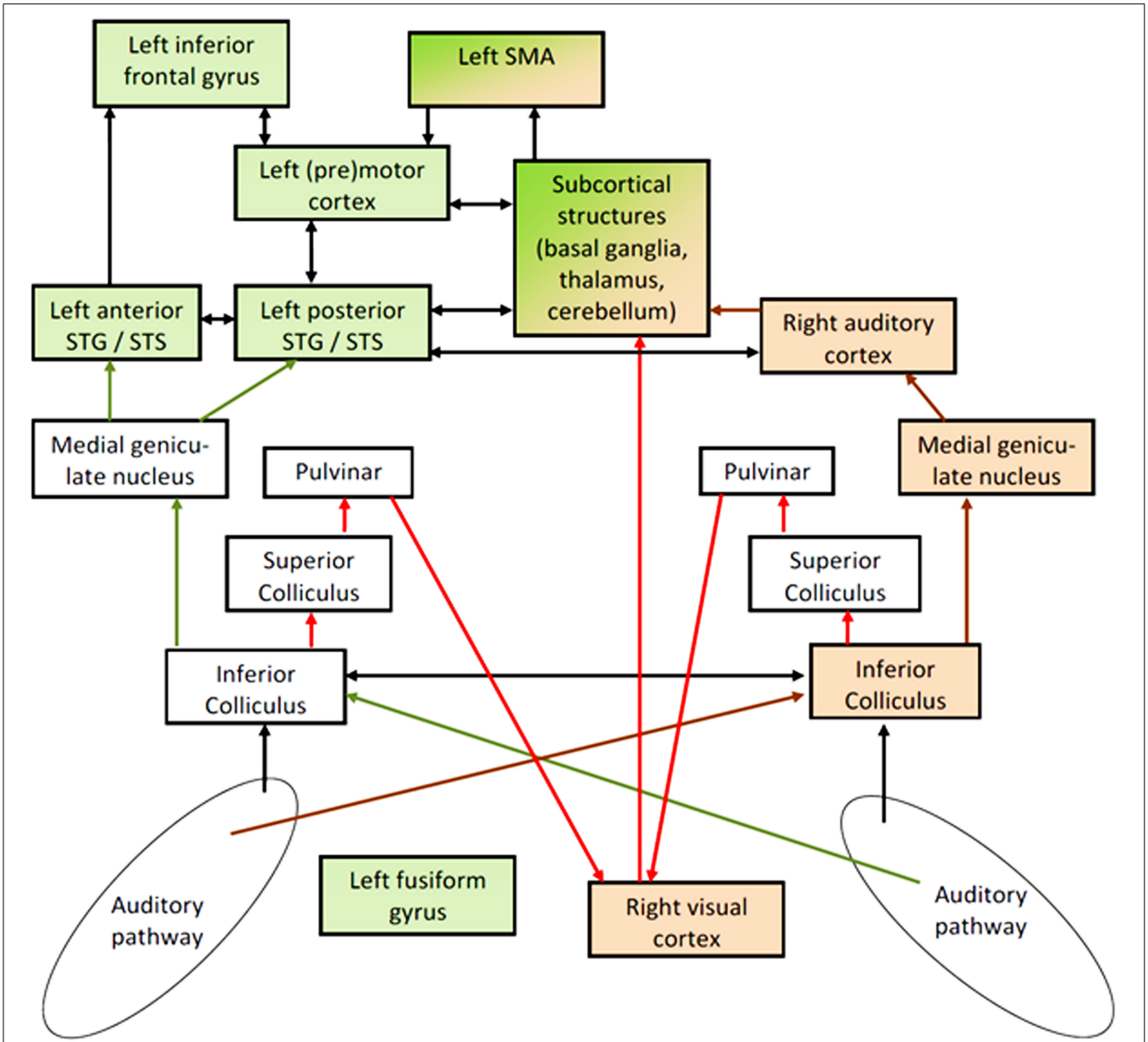


FIGURE 2 | Hypothetical pathways of speech perception: the phonological network – including secondary areas of left-hemisphere auditory cortex in superior temporal gyrus and sulcus (STG/STS) and frontal speech generation mechanisms – is colored in green, including additionally left fusiform gyrus (FG) in blind subjects. This network seems to be linked to a right-dominant syllable-prosodic network via subcortical structures and supplementary motor area (SMA). In normal subjects, this prosodic network is mainly localized in the right-hemisphere auditory system (brown arrows). In order to overcome temporal constraints regarding this prosodic stream as an independent signal (independent from segmental processing and from pitch processing), blind subjects seem to be able to recruit part of their visual cortex – presumably via subcortical afferent auditory information (red arrows) – to represent this prosodic information and to transfer it as an event-trigger channel to the frontal part of the speech processing network. Arrows to and from left FG were omitted in order to avoid an overload of the model and since the major aspect addressed here is the interplay between the right-dominant prosodic and the left-dominant phonological network. Furthermore, direct pathways between visual and auditory cortex were also omitted since the "bottleneck" for understanding ultra-fast speech seems to be located in the interface between sensory processing and frontal speech generation mechanisms.

2002; Noppeney, 2007). As a phylogenetic example, blind mole rats, rodents with a largely inactive peripheral visual system, have developed an additional pathway conveying auditory input from inferior colliculus via dorsal lateral geniculate nucleus to the central visual system (Bronchti et al., 2002). In humans, however, this connection between the afferent auditory and the primary visual pathway does not seem to be implemented.

Our recent studies on blind subjects point to a further possibility of visual cortex involvement in an auditory task, i.e., listening to time-compressed speech. As a substitute for reading, blind

individuals often use text-to-speech systems for the reception of texts. The speaking rate of these systems can be adjusted to quite high syllable rates, and blind users of these systems may learn to comprehend speech at rates up to ca. 20 syllables/s (Moos and Trouvain, 2007) while the normal speaking rate amounts to only 4–8 syllables/s. fMRI in blind subjects with the ability to understand ultra-fast speech at 16 syllables/s has shown hemodynamic activation, first, in left FG, a region that might be related to phonological representations (Cone et al., 2008) and, second, in right primary and secondary visual cortex, including parts of Brodmann areas (BA) 17 and 18 (Hertrich et al., 2009; Dietrich et al., 2013). Covariance analysis of fMRI data, furthermore, showed the ability to comprehend ultra-fast speech to be significantly associated, in addition to these two visual cortex areas, with activation in bilateral Pv, left IFG, left premotor cortex, left SMA as well as left anterior (aSTS) and bilateral posterior superior temporal sulcus (pSTS). As indicated by preliminary dynamic causal modeling (DCM) analyzes correlating functional connectivity with behavioral performance (Dietrich et al., 2010, 2011), the two visual areas activated in blind subjects, i.e., left-hemisphere FG and right-hemisphere primary and secondary visual cortex, seem to belong to different networks since they did not show significant connectivity in this analysis. FG, as part of the object-related ventral visual pathway (Haxby et al., 1991, 2000), might serve the representation of phonological "objects" linked to auditory and visual word form representations of the mental lexicon (McCandliss et al., 2003; Vigneau et al., 2006). Direct links between auditory and visual object representations have also been suggested to be activated by the use of sensory substitution devices "translating" optical signals into audible acoustic patterns (Striem-Amit et al., 2012). By contrast, right-dominant activation of early visual cortex as documented by Dietrich et al. (2013) seems to be associated with more elementary signal-related aspects as indicated by functional connectivity to pulvinar and auditory cortex. Furthermore, significant connectivity was observed between right visual cortex and left SMA, an area of temporal coordination in the frontal action network. Admittedly, considering the low temporal resolution of fMRI, this DCM analysis does not directly reflect the rapid information flow during speech perception. However, further evidence for an early signal-related rather than a higher-order linguistic aspect of speech processing being performed in right visual cortex has been provided by an MEG experiment (Hertrich et al., 2013). This study showed a particular signal component with a magnetic source in right occipital cortex that is phase-locked to a syllable onset signal derived from the speech envelope. The cross-correlation latency of this component was about 40–80 ms (see Figure 3 in Hertrich et al., 2013), indicating that this phase-locked activity arises quite early and, thus, might be driven by subcortical afferent input rather than cortico-cortical pathways. This might also be taken as an indicator that visual cortex activity represents a timing pattern rather than linguistic content. Thus, we hypothesize that visual cortex transfers a pre-linguistic prosodic signal, supporting the frontal action part of the speech perception network with timing information if the syllable rate exceeds the temporal resolution of the normal auditory prosody module. Admittedly, this model is still highly speculative given the limited basis of experimental data available so far. In addition, however, these suggestions shed some further light on exceptional abilities of blind subjects in the non-speech domain such as their resistance to backward masking as indicated by psychoacoustic experiments, pointing to a general mechanism of visual cortex recruitment for the purpose of time-critical event recording in blind subjects.

Taken together, left- and right-hemisphere activities observed in visual cortex of blind subjects during ultra-fast speech perception seem to be bound to the segmental (left) and prosodic (right) aspects of speech processing, in analogy to the Asymmetric Sampling hypothesis of the auditory system (Poeppel, 2003; Hickok and Poeppel, 2007). Activations of left-hemisphere phonological areas in the ventral visual stream can largely be expected on the basis of our knowledge regarding phonological and visual word form representations. By contrast, right visual cortex in blind subjects seems to belong to a different subsystem, receiving an afferent auditory timing signal that is related to syllable onsets and serving a similar function as the right-dominant prosodic timing channel in the theta band postulated for the auditory system (Abrams et al., 2008; Luo and Poeppel, 2012). However, the "prosodic" interpretation of right-hemisphere visual activities may require further support, first, with respect to existing pathways that could be able to build up such an extended prosodic network and, second, with respect to temporal resolution. Thus, in the following section various audiovisual experiments will be reviewed that can shed some light on the pathways contributing to visual system involvement in syllabic prosody representations.

AUDIOVISUAL EFFECTS AND ASSOCIATED PATHWAYS

Very robust perceptual audiovisual interactions have been documented, such as the sound-induced multiple flash illusion. Irrespective of spatial disparity, these experiments have demonstrated that visual perception can be qualitatively altered by auditory input at an early level of processing. In case of this illusion, for example, a (physical) single flash is perceived as a double-flash if it is accompanied by a sequence of two short acoustic signals (Shams et al., 2000; Shams and Kim, 2010). The perception of the illusory second flash has been found to depend upon an early electrophysiological response component in the central visual system following the second sound at a latency of only 30–60 ms (Mishra et al., 2007). These experiments nicely show that the visual cortex is well able to capture acoustic event information at a high temporal resolution and at an early stage of processing. Further electrophysiological evidence for very fast audiovisual interactions has been obtained during simple reaction time tasks (Molholm et al., 2002).

Under natural conditions, early auditory-to-visual information transfer may serve to improve the detection of visual events although it seems to work in a quite unspecific manner with respect to both the location of the visual event in the visual field and cross-modal spatial congruence or incongruence (Fiebelkorn et al., 2011). Furthermore, spatially irrelevant sounds presented shortly before visual targets may speed up reaction times, even in the absence of any specific predictive value (Keetels and Vroomen, 2011). Such early audio-to-visual interactions seem to work predominantly as timing cues rather than signaling specific event-related attributes although some auditory spatial information can, in addition, be derived, e.g., when two

data streams have to be segregated (Heron et al., 2012). Interestingly, the enhancement of visual target detection by auditory-to-visual information flow is not restricted to the actual event. Even passive repetitive auditory stimulation up to 30 min prior to a visual detection task can improve flash detection in the impaired hemifield of hemianopic patients (Lewald et al., 2012), indicating that auditory stimuli activate audiovisual pathways.

From a more general functional point of view, early audiovisual interactions facilitate the detection of cross-modal (in-)coherence of signals extending across both modalities. In this respect, there seems to be an asymmetry between the two channels with respect to temporal and spatial processing. In the temporal domain, the visual system appears to be adapted or gated (Purushothaman et al., 2012) by auditory information related to the time of acoustic signal onset (auditory dominance for timing). As a second step, the spatial representation of events within the dorsal auditory pathway may become recalibrated by coincident visual information (Wozny and Shams, 2011; spatial dominance of the visual system). This asymmetry, attributing temporal and spatial recalibration to different processing stages, can elucidate, for example, the differential interactions of these signal dimensions during the McGurk phenomenon (visual influence on auditory phonetic perception) as compared to the ventriloquist effect (visually induced spatial assignment of a speech signal to a speaking puppet Bishop and Miller, 2011). The McGurk effect is highly resistant against spatial incongruence, indicating an early binding mechanism (prior to the evaluation of spatial incongruence) on the basis of approximate temporal coincidence, followed by higher-order transfer of visual phonetic cues toward the auditory phonetic system. The temporal integration window of this effect has an asymmetrical structure and requires, as in natural stop consonant production, a temporal lag of the acoustic relative to the visual signal (Van Wassenhove et al., 2007). In this case, the visual component of the McGurk stimuli not only modifies, but also accelerates distinct electrophysiological responses such as the auditory-evoked N1 deflection (Van Wassenhove et al., 2005). However, an apparent motion design in which the shift between two pictures is exactly adjusted to the acoustic signal onset does not show such a visual effect on the auditory N1 response (Miki et al., 2004). In this latter case, presumably, early binding is not possible since the acoustic event trigger precedes the visual shift because of the delayed processing of actual visual signals. Thus, the McGurk effect seems to be based on a very early auditory-to-visual binding mechanism although its outcome might be the result of later higher-order phonological operations. By contrast, in case of the ventriloquist effect, the binding can be attributed to a later stage of spatial recalibration, top-down-driven by the perception of meaningful visual speech cues.

In contrast to syllabic event timing mechanisms assumed to engage visual cortex during ultra-fast speech perception, visuospatial cues are more or less irrelevant for blind subjects. The short latency (40–80 ms) of the MEG signal component phase-locked to syllable onsets over the right visual cortex (Hertrich et al., 2013) is comparable to the latency of visual cortex activity in case of the illusory double-flash perception, indicating a very early rather than late mechanism of visual cortex activation. As a consequence, we hypothesize that auditory timing information is derived from the acoustic signal at a pre-cortical stage, presumably, at the level of the SC, and then transferred to visual cortex via pulvinar and the posterior part of the secondary visual pathway. Although this pathway has been reported to target higher rather than primary visual areas (Martin, 2002; Berman and Wurtz, 2008, 2011), a diffusion tensor imaging tractography study indicates also the presence of connections from pulvinar to early cortical visual regions (Leh et al., 2008). As indicated by a monkey study, the pathway from pulvinar to V1 has a powerful gating function on visual cortex activity (Purushothaman et al., 2012). In sighted human subjects, the pulvinar-cortical visual pathway seems to play an important role with respect to Redundant Signal Effects (Maravita et al., 2008; see also Miller (1982) for behavioral effects of bimodal redundancy), multisensory spatial integration (Leo et al., 2008), audiovisual training of oculomotor functions during visual exploration (Passamonti et al., 2009), and suppression of visual motion effects during saccades (Berman and Wurtz, 2008, 2011). Regarding audiovisual interactions in sighted subjects such as the auditory-induced double-flash illusion (Shams et al., 2000; Mishra et al., 2007), the short latencies of electrophysiological responses of only 30–60 ms, by and large, rule out any significant impact of higher-order pathways from supramodal cortical regions to primary and secondary visual cortex as potential sources of this phenomenon, and even cross-modal cortico-cortical interactions between primary auditory and visual cortex might by too slow.

Cross-modal gating functions at the level of the auditory evoked P50, N100/M100 potentials as well as mismatch responses could be demonstrated within the framework of visual-to-auditory processing (Lebib et al., 2003; Van Wassenhove et al., 2005; Hertrich et al., 2007, 2009, 2011). Given that auditory event detection triggers visual event perception as in case of the auditory-induced double-flash illusion, it also seems possible that subcortical auditory information can trigger "visual" dummy events in the visual cortex of blind subjects. Subsequently, these event markers may function as a secondary temporal gating signal for the purpose of phonological encoding.

Frontal cortex, particularly, SMA, seems to play an important role in the coordination of phonological encoding with prosodic timing (see above). In principle, visual and audiovisual information via SC and pulvinar might reach frontal cortex in the absence of any activation of the occipital lobe (Liddell et al., 2005). However, this pathway is unlikely to be involved in the perception of ultra-fast speech since, first, it does not particularly involve SMA and, second, it is linked to reflexive action rather than conscious perception. Thus, we assume that in order to signalize an event-related trigger signal to the SMA, the data stream has to pass sensory cortical areas such somatosensory, auditory, or visual cortex. But how can audiovisual events (in sighted) or auditory-induced empty events represented in visual cortex (in blind people) feed timing information into SMA? A comprehensive study of the efferent and afferent connections of this mesiofrontal area in squirrel monkeys found multiple cortical and subcortical pathways, but no direct input from primary or secondary visual cortex. By contrast, proprioception, probably due to its close relationship to motor control, seems to have a more direct influence on SMA activity (Jürgens, 1984). Regarding the visual domain, SMA seems to be involved in visually cued motor

tasks (Mohamed et al., 2003) and in visually guided tracking tasks (Picard and Strick, 2003) as well as in an interaction of visual event detection with oral conversation as shown by reaction time effects (Bowyer et al., 2009). Thus, in analogy to the auditory models of Hickok and Poeppel (2007) and Kotz and Schwartze (2010), we may assume a pathway from the right-hemisphere dorsal visual stream, representing syllabic events, toward the SMA via subcortical structures including the thalamus and the (left) cerebellum.

DISCUSSION

In summary, the present model assumes a dual data stream to support the linguistic encoding of continuous speech: predominant left-hemisphere extraction of phonetic features and predominant right-hemisphere capture of the speech envelope. The coordination of these two functional subsystems seems to be bound to the frontal cortex. More specifically, SMA might critically contribute to the synchronization of the incoming signal with top-down driven syllabically organized sequential pacing signals. In case of ultra-fast speech, the auditory system – although capable to process signals within the 16 Hz domain – may fail to separate syllable-prosodic and segmental information at such high rates. Therefore, the speech generation system, including the phonological working memory, cannot be triggered by a prosodic event channel. In order to overcome this bottleneck, we must either learn to encode speech signals in the absence of a syllabic channel – a, most presumably, quite difficult task – or we have to recruit a further neural pathway to provide the frontal cortex with syllabic information. The latter strategy seems to be available to blind subjects who may use the audiovisual interface of the secondary visual pathway in order to transmit syllabic event triggers via pulvinar to right visual cortex. As a consequence, the tentative function of visual cortex might consist in the transformation of the received timing signal into a series of (syllabic) events that subsequently can be conveyed to the frontal lobe in order to trigger the phonological representations in the speech generation and working memory system. These "events" might be similar to the ones that, in sighted subjects, become spatially recalibrated by vision. Since vision loss precludes any spatial recalibration, the auditory events may target a region near the center of the retinotopic area in visual cortex. Considering, first, that this audiovisual pathway is linked to visuospatial processing in sighted subjects and, second, that the extracted auditory signal components are prosodic event-related rather than phonological data structures, it seems rather natural that they are preferably processed within the right-hemisphere. Thus, by "outsourcing" the syllabic channel into the visual system, blind people may overcome the prosodic event timing limits of right-hemisphere auditory cortex.

Various aspects of the proposed model must now be tested explicitly, e.g., by means of TMS techniques and further connectivity analyzes. Assuming, for example, that right visual cortex of blind subjects is involved in prosodic timing mechanisms, a virtual lesion of this area during ultra-fast speech perception must be expected to yield similar comprehension deficits as virtual damage to right auditory cortex in sighted subjects during perception of moderately fast speech. Furthermore, pre-activation of right visual cortex as well as co-activation of right visual cortex with SMA might have facilitating effects on speech processing. In sighted subjects, furthermore, it should be possible to simulate the early phase-locked activity in right visual cortex by presenting flashes that are synchronized with syllable rate. If, indeed, visual cortex can forward prosodic event triggers, these flashes should enhance the comprehension of time-compressed speech.

So far, only few studies provide clear-cut evidence for a subcortical audiovisual pathway targeting primary visual cortex. The present model postulates that a speech envelope signal is already represented at a pre-cortical level of the brain. As a consequence, the prosodic timing channel engaged in speech processing should be separated from the "segmental" auditory channel already at a subcortical stage. So far, recordings of brainstem potentials did not reveal any lateralization effects similar to the cortical distinction of short-term segmental (left hemisphere) and low-frequency suprasegmental/prosodic (right-hemisphere) information (Abrams et al., 2010). At the level of the thalamus, however, low-frequency information is well represented, and it has been hypothesized that these signals – bound predominantly to paralemniscal pathways – have a gating function regarding the perceptual evaluation of auditory events (He, 2003; Abrams et al., 2011). Furthermore, the underlying temporal coding mechanism (spike timing) seems to be particularly involved in the processing of communication sounds via thalamus, primary and non-primary auditory cortex up to frontal areas (Huetz et al., 2011).

Alternatively, one might suggest that the visual cortex of blind individuals is activated by cross-modal cortico-cortical pathways. In sighted subjects, however, early audiovisual interactions allowing for the enhancement of auditory processing by visual cues require a time-lead of the visual channel extending from 20 to 80 ms (Kayser et al., 2008). Thus, it seems implausible that ultra-fast speech comprehension can be accelerated by visual cortex activation via cortico-cortical cross-modal pathways. If the visual channel is really capable to impact auditory encoding of speech signals at an early phase-locked stage, then very early subcortical afferent input to the visual system must be postulated. These fast connections might trigger phonological encoding in a manner analogous to the prosodic timing mechanisms in right-hemisphere auditory cortex. The underlying mechanism of this process might consist in phase modulation of oscillatory activity within visual cortex based on subcortical representations of the speech envelope.

Since the "bottleneck" for understanding ultra-fast speech in sighted subjects has been assigned to frontal rather than temporal regions, pathways projecting from visual to frontal cortex, targeting, in particular, SMA, must be assumed in order to understand how blind people can overcome these constraints. The connections sighted subjects use to control the motor system during visual perception, both in association with ocular and visually guided upper limb movements, represent a plausible candidate structure. Considering SMA a motor timing device with multiple input channels but no direct interconnections with primary visual cortex, the transfer of the prosodic signals toward SMA might be performed via subcortical mechanisms involving cerebellum, basal ganglia, and thalamus. However, in upcoming studies this has to be demonstrated explicitly.

The present model might also contribute to a better understanding of previous findings on enhanced auditory performance of blind individuals such as resistance to backward masking, as documented by Stevens and Weaver (2005). Thereby, this aspect of temporal processing seems to be related to perceptual consolidation rather than elementary auditory time resolution. Furthermore, resistance to backward masking in blind subjects was associated with activity, even preparatory activity in visual cortex. In line with the present model, activation of visual cortex was found in the right rather than the left hemisphere. Stevens et al. (2007) interpreted the preparatory visual activation as a "baseline shift" related to attentional modulation. However, they did not provide an explicit hypothesis about the nature of the input signal toward visual cortex. Based on the present model, we might assume that the secondary visual pathway provides the visual system with afferent auditory information. Considering brain activations outside the visual system, Stevens et al. (2007) did not mention SMA, but other frontal regions such as the frontal eye field, known as a structure serving auditory attentional processing in blind subjects (Garg et al., 2007). Thus, at least some aspects of the present model might be expanded to the non-speech domain, referring to a general mechanism that enhances the temporal resolution of auditory event recording by using the afferent audiovisual interface toward the secondary visual pathway.

At least partially, the assumption of an early signal-related transfer mechanism via pulvinar, secondary visual pathway, and right visual cortex toward the frontal cortex was based on fMRI connectivity analyzes, an approach of still limited temporal resolution. So far, it cannot be excluded that frontal cortex activation under these conditions simply might reflect higher-order linguistic processes that are secondary to, but not necessary for comprehension. Nevertheless, functional imaging data revealed the time constraints of speech understanding to be associated with frontal structures (Vagharchakian et al., 2012). Thus, frontal lobe activity during spoken language comprehension seems comprise both the generation of inner speech after lexical access and the generation of well-timed predictions regarding the syllabically organized structure of upcoming speech material. In other words, it is an interface between bottom-up and top-down mechanisms.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG; AC 55 09/01) and the Hertie Institute for Clinical Brain Research, Tübingen. Furthermore, we acknowledge support by the Open Access Publishing Fund of the University of Tuebingen, sponsored by the DFG.

REFERENCES

Abler, B., Roebroeck, A., Goebel, R., Höse, A., Schönfeldt-Lecuona, C., Hole, G., et al. (2006). Investigating directed influences between activated brain areas in a motor-response task using fMRI. *Magn. Reson. Imaging* 24, 181–185. doi: 10.1016/j.mri.2005.10.022

Abrams, D. A., Nicol, T., Zecker, S., and Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *J. Neurosci.* 28, 3958–3965. doi: 10.1523/JNEUROSCI.0187-08.2008

Abrams, D. A., Nicol, T., Zecker, S., and Kraus, N. (2010). Rapid acoustic processing in the auditory brainstem is not related to cortical asymmetry for the syllable rate of speech. *Clin. Neurophysiol.* 121, 1343–1350. doi: 10.1016/j.clinph.2010.02.158

Abrams, D. A., Nicol, T., Zecker, S., and Kraus, N. (2011). A possible role for a paralemniscal auditory pathway in the coding of slow temporal information. *Hear. Res.* 272, 125–134. doi: 10.1016/j.heares.2010.10.009

Acheson, D. J., Hamidi, M., Binder, J. R., and Postle, B. R. (2010). A common neural substrate for language production and verbal working memory. *J. Cogn. Neurosci.* 23, 1358–1367. doi: 10.1162/jocn.2010.21519

Amedi, A., Floel, A., Knecht, S., Zohary, E., and Cohen, L. G. (2004). Transcranial magnetic stimulation of the occipital pole interferes with verbal processing in blind subjects. *Nat. Neurosci.* 7, 1266–1270. doi: 10.1038/nn1328

Bavelier, D., and Hirshorn, E. A. (2010). I see where you're hearing: how cross-modal plasticity may exploit homologous brain structures. *Nat. Neurosci.* 13, 1309–1311. doi: 10.1038/nn1110-1309

Bavelier, D., and Neville, H. J. (2002). Cross-modal plasticity: where and how? *Nat. Rev. Neurosci.* 3, 443–452. doi: 10.1038/nrn848

Berman, R. A., and Wurtz, R. H. (2008). Exploring the pulvinar path to visual cortex. *Prog. Brain Res.* 171, 467–473. doi: 10.1016/S0079-6123(08)00668-7

Berman, R. A., and Wurtz, R. H. (2011). Signals conveyed in the pulvinar pathway from superior colliculus to cortical area MT. *J. Neurosci.* 31, 373–384. doi: 10.1523/JNEUROSCI.4738-10.2011

Bishop, C. W., and Miller, L. M. (2011). Speech cues contribute to audiovisual spatial integration. *PLoS ONE* 6:e24016. doi: 10.1371/journal.pone.0024016

Bowyer, S. M., Hsieh, L., Moran, J. E., Young, R. A., Manoharan, A., Liao, C.-C. J., et al. (2009). Conversation effects on neural mechanisms underlying reaction time to visual events while viewing a driving scene using MEG. *Brain Res.* 1251, 151–161. doi: 10.1016/j.brainres.2008.10.001

Brendel, B., Hertrich, I., Erb, M., Lindner, A., Riecker, A., Grodd, W., et al. (2010). The contribution of mesiofrontal cortex to the preparation and execution of repetitive syllable productions: an fMRI study. *Neuroimage* 50, 1219–1230. doi: 10.1016/j.neuroimage.2010.01.039

Bronchti, G., Heil, P., Sadka, R., Hess, A., Scheich, H., and Wollberg, Z. (2002). Auditory activation of "visual" cortical areas in the blind mole rat (*Spalax ehrenbergi*). *Eur. J. Neurosci.* 16, 311–329. doi: 10.1046/j.1460-9568.2002.02063.x

Brugge, J. F., Nourski, K. V., Oya, H., Reale, R. A., Kawasaki, H., Steinschneider, M., et al. (2009). Coding of repetitive transients by auditory cortex on Heschl's gyrus. *J. Neurophysiol.* 102, 2358–2374. doi: 10.1152/jn.91346.2008

Buccino, G., Binkofski, F., and Riggio, L. (2004). The mirror neuron system and action recognition. *Brain Lang.* 89, 370–376. doi: 10.1016/S0093-934X(03)00356-0

Büchel, C. (2003). Cortical hierarchy turned on its head. *Nat. Neurosci.* 6, 657–658. doi: 10.1038/nn0703-657

Büchel, C., Price, C., Frackowiak, R. S. J., and Friston, K. (1998). Different activation patterns in the visual cortex of late and congenitally blind subjects. *Brain* 121, 409–419. doi: 10.1093/brain/121.3.409

Buchsbaum, B. R., and D'Esposito, M. (2008). The search for the phonological store: from loop to convolution. *J. Cogn. Neurosci.* 20, 762–778. doi: 10.1162/jocn.2008.20501

Burton, H. (2003). Visual cortex activity in early and late blind people. *J. Neurosci.* 23, 4005–4011.

Burton, H., Sinclair, R. J., and Dixit, S. (2010). Working memory for vibrotactile frequencies: comparison of cortical activity in blind and sighted individuals. *Hum. Brain Mapp.* 31, 1686–1701. doi: 10.1002/hbm.20966

Cao, F., Bitan, T., and Booth, J. R. (2008). Effective brain connectivity in children with reading difficulties during phonological processing. *Brain Lang.* 107, 91–101. doi: 10.1016/j.bandl.2007.12.009

Cone, N. E., Burman, D. D., Bitan, T., Bolger, D. J., and Booth, J. R. (2008). Developmental changes in brain regions involved in phonological and orthographic processing during spoken language processing. *Neuroimage* 41, 623–635. doi: 10.1016/j.neuroimage.2008.02.055

D'Ausilio, A., Craighero, L., and Fadiga, L. (2012). The contribution of the frontal lobe to the perception of speech. *J. Neuroling.* 25, 328–335. doi: 10.1016/j.jneuroling.2010.02.003

De Witt, I., and Rauschecker, J. P. (2012). Phoneme and word

recognition in the auditory ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* 109, E505–E514. doi: 10.1073/pnas.1113427109

Dietrich, S., Hertrich, I., and Ackermann, H. (2010). Visual cortex doing an auditory job: enhanced spoken language comprehension in blind subjects. *Abstract Society for Neuroscience 2010.* Available at: http://www.abstractsonline.com/Plan/ViewAbstract.aspx? (accessed February 1, 2013).

Dietrich, S., Hertrich, I., and Ackermann, H. (2011). Why do blind listeners use visual cortex for understanding ultra-fast speech? *ASA Lay Language Papers, 161st Acoustical Society of America Meeting 2011*. Available at: http://www.acoustics.org/press/161st/Dietrich.html (accessed February 1, 2013).

Dietrich, S., Hertrich, I., and Ackermann, H. (2013). Ultra-fast speech comprehension in blind subjects engages primary visual cortex, fusiform gyrus, and pulvinar – a functional magnetic resonance imaging (fMRI) study. *BMC Neurosci.* 14:74. doi: 10.1186/1471-2202-14-74.

Elbert, T., Sterr, A., Rockstroh, B., Pantev, C., ller, M. M., and Taub, E. (2002). Expansion of the tonotopic area in the auditory cortex of the blind. *J. Neurosci.* 22, 9941–9944.

Fiebelkorn, I. C., Foxe, J. J., Butler, J. S., and Molholm, S. (2011). Auditory facilitation of visual-target detection persists regardless of retinal eccentricity and despite wide audiovisual misalignments. *Exp. Brain Res.* 213, 167–174. doi: 10.1007/s00221-011-2670-7

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

Garg, A., Schwartz, D., and Stevens, A. A. (2007). Orienting auditory spatial attention engages frontal eye fields and medial occipital cortex in congenitally blind humans. *Neuropsychologia* 45, 2307–2321. doi: 10.1016/j.neuropsychologia.2007.02.015

Ghitza, O., and Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* 66, 113–126. doi: 10.1159/000208934

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. (2012). The cortical organization of lexical knowledge: a dual lexicon model of spoken language processing. *Brain Lang.* 121, 273–288. doi: 10.1016/j.bandl.2012.03.005

Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *J. Phon.* 31, 465–485. doi: 10.1016/j.wocn.2003.09.005

Grimaldi, M. (2012). Toward a neural theory of language: old issues and new perspectives. *J. Neurolinguistics* 25, 304–327. doi: 10.1016/j.jneuroling.2011.12.002

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., et al. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc. Natl. Acad. Sci. U.S.A.* 88, 1621–1625. doi: 10.1073/pnas.88.5.1621

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends Cogn. Sci.* 4, 223–233. doi: 10.1016/S1364-6613(00)01482-0

He, J. (2003). Slow oscillation in nonlemniscal auditory thalamus. *J. Neurosci.* 23, 8281–8290.

Heron, J., Roach, N. W., Hanson, J. V. M., McGraw, P. V., and Whitaker, D. (2012). Audiovisual time perception is spatially specific. *Exp. Brain Res.* 218, 477–485. doi: 10.1007/s00221-012-3038-3

Hertrich, I., Dietrich, S., and Ackermann, H. (2011). Cross-modal interactions during perception of audiovisual speech and non-speech signals: an fMRI study. *J. Cogn. Neurosci.* 23, 221–237. doi: 10.1162/jocn.2010.21421

Hertrich, I., Dietrich, S., and Ackermann, H. (2013). Tracking the speech signal – time-locked MEG signals during perception of ultra-fast and moderately fast speech in blind and in sighted listeners. *Brain Lang.* 124, 9–21. doi: 10.1016/j.bandl.2012.10.006

Hertrich, I., Dietrich, S., Moos, A., Trouvain, J., and Ackermann, H. (2009). Enhanced speech perception capabilities in a blind listener are associated with activation of fusiform gyrus and primary visual cortex. *Neurocase* 15, 163–170. doi: 10.1080/13554790802709054

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., and Ackermann, H. (2012). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology* 49, 322–334. doi: 10.1111/j.1469-8986.2011.01314.x

Hertrich, I., Mathiak, K., Lutzenberger, W., and Ackermann, H. (2004). Transient and phase-locked evoked magnetic fields in response to periodic acoustic signals. *Neuroreport* 15, 1687–1690. doi: 10.1097/01.wnr.0000134930.04561.b2

Hertrich, I., Mathiak, K., Lutzenberger, W., and Ackermann, H. (2009). Time course of early audiovisual interactions during speech and non-speech central auditory processing: a magnetoencephalography study. *J. Cogn. Neurosci.* 21, 259–274. doi: 10.1162/jocn.2008.21019

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

Hickok, G., and Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends Cogn. Sci.* 4, 131–138. doi: 10.1016/S1364-6613(00)01463-7

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

Huetz, C., Gourévitch, B., and Edeline, J.-M. (2011). Neural codes in the thalamocortical auditory system: from artificial stimuli to communication sounds. *Hear. Res.* 271, 147–158. doi: 10.1016/j.heares.2010.01.010

Jürgens, U. (1984). The efferent and afferent connections of the supplementary motor area. *Brain Res.* 300, 63–81. doi: 10.1016/0006-8993(84)91341-6

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cereb. Cortex* 18, 1560–1574. doi: 10.1093/cercor/bhm187

Keetels, M., and Vroomen, J. (2011). Sound affects the speed of visual processing. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 699–708. doi: 10.1037/a0020564

Keysers, C., and Fadiga, L. (2008). The mirror neuron system: new frontiers. *Soc. Neurosci.* 3, 193–198. doi: 10.1080/17470910802408513

Kotz, S. A., D'Ausilio, A., Raettig, T., Begliomini, C., Craighero, L., Fabbri-Destro, M., et al. (2010). Lexicality drives audio-motor transformations in Broca's area. *Brain Lang.* 112, 3–11. doi: 10.1016/j.bandl.2009.07.008

Kotz, S. A., and Schwartze, M. (2010). Cortical speech processing unplugged: a timely subcortico-cortical framework. *Trends Cogn. Sci.* 14, 392–399. doi: 10.1016/j.tics.2010.06.005

Kotz, S. A., Schwartze, M., and Schmidt-Kassow, M. (2009). Non-motor basal ganglia functions: a review and proposal for a model of sensory predictability in auditory language perception. *Cortex* 45, 982–990. doi: 10.1016/j.cortex.2009.02.010

Kupers, R., Beaulieu-Lefebvre, M., Schneider, F. C., Kassuba, T., Paulson, O. B., Siebner, H. R., et al. (2011). Neural correlates of olfactory processing in congenital blindness. *Neuropsychologia* 49, 2037–2044. doi: 10.1016/j.neuropsychologia.2011.03.033

Lebib, R., Papo, D., De Bode, S., and Baudonniere, 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

Leh, S. E., Chakravarty, M. M., and Ptito, A. (2008). The connectivity of the human pulvinar: a diffusion tensor imaging tractography study. *Int. J. Biomed. Imaging* 2008, 789539. doi: 10.1155/2008/789539

Leo, F., Bertini, C., di Pellegrino, G., and Làdavas, E. (2008). Multisensory integration for orienting responses in humans requires the activation of the superior colliculus. *Exp. Brain Res.* 186, 67–77. doi: 10.1007/s00221-007-1204-9

Lewald, J., Tegenthoff, M., Peters, S., and Hausmann, M. (2012). Passive auditory stimulation improves vision in hemianopia. *PLoS ONE* 7:e31603. doi: 10.1371/journal.pone.0031603

Liddell, B. J., Brown, K. J., Kemp, A. H., Barton, M. J., Das, P., Peduto, A., et al. (2005). A direct brainstem–amygdala–cortical "alarm" system for subliminal signals of fear. *Neuroimage* 24, 235–243. doi: 10.1016/j.neuroimage.2004.08.016

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

Luo, H., and Poeppel, D. (2012). Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Front. Psychol.* 3:170. doi: 10.3389/fpsyg.2012.00170

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., and Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS ONE* 4:e4638. doi: 10.1371/journal.pone.0004638

Macaluso, E., and Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends Neurosci.* 28, 264–271. doi: 10.1016/j.tins.2005.03.008

Maravita, A., Bolognini, N., Bricolo, E., Marzi, C. A., and Savazzi, S. (2008). Is audiovisual integration subserved by the superior colliculus in humans? *Neuroreport*, 19, 271–275. doi: 10.1097/WNR.0b013e3282f4f04e

Martin, E. (2002). "Imaging of brain function during early human development," in *MRI of the Neonatal Brain*, ed. M. Rutherford, Chapter 18, E-book. Available at: http://www.mrineonatalbrain.com/ch04-18.php

McCandliss, B. D., Cohen, L., and Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn. Sci.* 7, 293–299. doi: 10.1016/S1364-6613(03)00134-7

Meltzoff, A. N., and Moore, M. K. (1997). Explaining facial imitation: a theoretical model. *Early Dev. Parenting* 6, 179–192. doi: 10.1002/(SICI)1099-0917(199709/12)6:3/4<179::AID-EDP157>3.0.CO;2-R

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

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

Mishra, J., Martinez, A., Sejnowski, T. J., and Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *J. Neurosci.* 27, 4120–4131. doi: 10.1523/JNEUROSCI.4912-06.2007

Mohamed, M. A., Yousem, D. M., Tekes, A., Browner, N. M., and Calhoun, V. D. (2003). Timing of cortical activation: a latency-resolved event-related functional MR imaging study. *Am. J. Neuroradiol.* 24, 1967–1974.

Molenberghs, P., Cunnington, R., and Mattingley, J. B. (2012). Brain regions with mirror properties: a meta-analysis of 125 human fMRI studies. *Neurosci. Biobehav. Rev.* 36, 341–349. doi: 10.1016/j.neubiorev.2011.07.004

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

Moos, A., and Trouvain, J. (2007). "Comprehension of ultra-fast speech – blind vs. "normally hearing" persons," in *Proceedings of the sixteenth International Congress of Phonetic Sciences*, eds J. Trouvain and W. J. Barry (Saarbrücken: University of Saarbrücken), 677–680. Available at: http://www.icphs2007.de/conference/Papers/1186/1186.pdf

Noppeney, U. (2007). The effects of visual deprivation on functional and structural organization of the human brain. *Neurosci. Biobehav. Rev.* 31, 1169–1180. doi: 10.1016/j.neubiorev.2007.04.012

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., et al. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *J. Neurosci.* 29, 15564–15574. doi: 10.1523/JNEUROSCI.3065-09.2009

Passamonti, C., Bertini, C., and Làdavas, E. (2009). Audio-visual stimulation improves oculomotor patterns in patients with hemianopia. *Neuropsychologia* 47, 546–555. doi: 10.1016/j.neuropsychologia.2008.10.008

Patel, A. D. (2003). Language, music, syntax, and the brain. *Nat. Neurosci.* 6, 674–681. doi: 10.1038/nn1082

Perez-Pereira, M., and Conti-Ramsden, G. (1999). *Language Development and Social Interaction in Blind Children.* Hove, UK: Psychology Press Ltd.

Picard, N., and Strick, P. L. (2003). Activation of the supplementary motor area (SMA) during performance of visually guided movements. *Cereb. Cortex* 13, 977–986. doi: 10.1093/cercor/13.9.977

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

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

Purushothaman, G., Marion, R., Li, K., and Casagrande, V. A. (2012). Gating and control of primary visual cortex by pulvinar. *Nat. Neurosci.* 15, 905–912. doi: 10.1038/nn.3106

Ricciardi, E., and Pietrini, P. (2011). New light from the dark: what blindness can teach us about brain function. *Curr. Opin. Neurol.* 24, 357–363. doi: 10.1097/WCO.0b013e328348bdbf

Riecker, A., Mathiak, K., Wildgruber, D., Erb, M., Hertrich, I., Grodd, W., et al. (2005). fMRI reveals two distinct cerebral networks subserving speech motor control. *Neurology* 64, 700–706. doi: 10.1212/01.WNL.0000152156.90779.89

Röder, B., Kramer, U. M., and Lange, K. (2007). Congenitally blind humans use different stimulus selection strategies in hearing: an ERP study of spatial and temporal attention. *Restor. Neurol. Neurosci.* 25, 311–322.

Röder, B., Stock, O., Bien, S., Neville, H., and Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *Eur. J. Neurosci.* 16, 930–936. doi: 10.1046/j.1460-9568.2002.02147.x

Romani, C., Galluzzi, C., and Olson, A. (2011). Phonological–lexical activation: a lexical component or an output buffer? Evidence from aphasic errors. *Cortex* 47, 217–235. doi: 10.1016/j.cortex.2009.11.004

Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions: what you see is what you hear. *Nature* 408, 788–788. doi: 10.1038/35048669

Shams, L., and Kim, R. (2010). Crossmodal influences on visual perception. *Phys. Life Rev.* 7, 269–284. doi: 10.1016/j.plrev.2010.04.006

Spence, M. J., and Decasper, A. J. (1987). Prenatal experience with low-frequency material-voice sounds influence neonatal perception of maternal voice samples. *Infant Behav. Dev.* 10, 133–142. doi: 10.1016/0163-6383(87)90028-2

Stevens, A. A., Snodgrass, M., Schwartz, D., and Weaver, K. (2007). Preparatory activity in occipital cortex in early blind humans predicts auditory perceptual performance. *J. Neurosci.* 27, 10734–10741. doi: 10.1523/JNEUROSCI.1669-07.2007

Stevens, A. A., and Weaver, K. (2005). Auditory perceptual consolidation in early-onset blindness. *Neuropsychologia* 43, 1901–1910. doi: 10.1016/j.neuropsychologia.2005.03.007

Stevens, A. A., and Weaver, K. E. (2009). Functional characteristics of auditory cortex in the blind. *Behav. Brain Res.* 196, 134–138. doi: 10.1016/j.bbr.2008.07.041

Streri, A., Coulon, M., and Guellaï, B. (2013). The foundations of social cognition: studies on face/voice integration in newborn infants. *Int. J. Behav. Dev.* 37, 79–83. doi: 10.1177/0165025412465361

Striem-Amit, E., Guendelman, M., and Amedi, A. (2012). "Visual" acuity of the congenitally blind using visual-to-auditory sensory substitution. *PLoS ONE* 7:e33136. doi: 10.1371/journal.pone.0033136

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

Tanji, J. (1994). The supplementary motor area in the cerebral cortex. *Neurosci. Res.* 19, 251–268. doi: 10.1016/0168-0102(94)90038-8

Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., and Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *J. Neurosci.* 32, 9089–9102. doi: 10.1523/JNEUROSCI.5685-11.2012

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

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

Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., et al. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage* 30, 1414–1432. doi: 10.1016/j.neuroimage.2005.11.002

Wozny, D. R., and Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. *J. Neurosci.* 31, 4607–4612. doi: 10.1523/JNEUROSCI.6079-10.2011

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 February 2013; accepted: 26 July 2013; published online: 16 August 2013.
*Citation: Hertrich I, Dietrich S and Ackermann H (2013) How can audio-visual pathways enhance the temporal resolution of time-compressed speech in blind subjects? Front. Psychol. **4**:530. doi: 10.3389/fpsyg.2013.00530*
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 June 2013
doi: 10.3389/fpsyg.2013.00331

Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs

Sanne ten Oever[1]*, Alexander T. Sack[1], Katherine L. Wheat[1], Nina Bien[1,2] and Nienke van Atteveldt[1,3]

[1] Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands
[2] EMACS Research Unit, University of Luxembourg, Luxembourg, Luxembourg
[3] Neuroimaging and Neuromodeling Group, Netherlands Institute for Neuroscience, Amsterdam, Netherlands

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
Adele Diederich, Jacobs University Bremen, Germany
Argiro Vatakis, Cognitive Systems Research Institute, Greece
Ryan A. Stevenson, Vanderbilt University Medical Center, USA

***Correspondence:**
Sanne ten Oever, Faculty of Psychology and Neuroscience, Maastricht University, Oxfordlaan 55, 6200 MD Maastricht, Netherlands
e-mail: sanne.tenoever@maastrichtuniversity.nl

Content and temporal cues have been shown to interact during audio-visual (AV) speech identification. Typically, the most reliable unimodal cue is used more strongly to identify specific speech features; however, visual cues are only used if the AV stimuli are presented within a certain temporal window of integration (TWI). This suggests that temporal cues denote whether unimodal stimuli belong together, that is, whether they should be integrated. It is not known whether temporal cues also provide information about the identity of a syllable. Since spoken syllables have naturally varying AV onset asynchronies, we hypothesize that for suboptimal AV cues presented within the TWI, information about the natural AV onset differences can aid in speech identification. To test this, we presented low-intensity auditory syllables concurrently with visual speech signals, and varied the stimulus onset asynchronies (SOA) of the AV pair, while participants were instructed to identify the auditory syllables. We revealed that specific speech features (e.g., voicing) were identified by relying primarily on one modality (e.g., auditory). Additionally, we showed a wide window in which visual information influenced auditory perception, that seemed even wider for congruent stimulus pairs. Finally, we found a specific response pattern across the SOA range for syllables that were not reliably identified by the unimodal cues, which we explained as the result of the use of natural onset differences between AV speech signals. This indicates that temporal cues not only provide information about the temporal integration of AV stimuli, but additionally convey information about the identity of AV pairs. These results provide a detailed behavioral basis for further neuro-imaging and stimulation studies to unravel the neurofunctional mechanisms of the audio-visual-temporal interplay within speech perception.

Keywords: audiovisual, temporal cues, audio-visual onset differences, content cues, predictability, detection

INTRODUCTION

Although audition is our main informant during speech perception, visual cues have been shown to strongly influence identification and recognition of speech (Campbell, 2008). Visual cues are used to increase understanding, especially in noisy situations when auditory information alone is not sufficient (Sumby and Pollack, 1954; Bernstein et al., 2004; Grant et al., 2004). It is known that temporal, spatial, and semantic cues in visual signals are used to improve auditory speech perception (Wallace et al., 1996; Stevenson and James, 2009). However, it is largely unknown how these different cues are combined to create our auditory percept. In the current research, we used semantically congruent or incongruent audio-visual syllables presented with varied stimulus onset asynchronies (SOAs) between the auditory and visual stimuli, to investigate the interaction between temporal and content factors during audio-visual speech perception (see e.g., Vatakis and Spence, 2006; van Wassenhove et al., 2007; Vatakis et al., 2012). Specifically, we were interested whether natural onset asynchronies inherent to audio-visual syllable pairs influence syllable identification.

Often, stop-consonant syllables (e.g., /ba/ and /da/) are used to examine syllable identification (see e.g., McGurk and MacDonald, 1976; van Wassenhove et al., 2007; Arnal et al., 2011). Stop consonants are consistent in the manner in which they are produced (the vocal tract is blocked to cease airflow), but vary in the type and amount of identity information conveyed by the visual and auditory channels. Specifically, whether or not the vocal tract is used to produce a consonant (i.e., the voicing of a sound, /ba/ vs. /pa/) is not visible, since the vocal tract is located in the throat. Therefore, the auditory signal is more reliable than the visual signal in determining the voicing of a speech signal (Wiener and Miller, 1946; McGurk and MacDonald, 1976). On the other hand, which part of the mouth we use for producing a syllable is mostly a visual signal. For example, uttering a syllable with our lips (like /ba/) vs. our tongue (like /da/) is more visible than audible. Visual speech thus conveys mostly information about the place of articulation (POA) of the sound, and adding acoustic noise to a spoken syllable makes the POA particularly difficult to extract on basis of auditory information (Wiener and Miller, 1946; McGurk and MacDonald, 1976; van Wassenhove et al., 2005). However, the amount of visual

information about the POA varies for different syllables: bilabial syllables (pronounced with the lips) are better dissociated than coronal and dorsal syllables (pronounced with the front or body of the tongue, respectively). Thus, it seems that auditory and visual speech signals are complementary in identifying a syllable, since voicing information is best conveyed by auditory cues and POA information by visual cues (Summerfield, 1987; Campbell, 2008).

Auditory and visual stimuli can be linked based on their content information; the information about the identity (the "what") of a stimulus. We will continue to use the term content information, although in other studies the term semantic information is also used (for a review, see Doehrmann and Naumer, 2008). The amount of content information conveyed by a unimodal signal is variable, for different stimuli (as explained above) as well as for individuals perceiving the same stimuli, and the reliability of the information determines how strongly it influences our percept (Driver, 1996; Beauchamp et al., 2004; van Wassenhove et al., 2005; Blau et al., 2008). For example, the amount of content information present in visual speech signals is widely variable, as reflected in individual differences in lipreading skills (MacLeod and Summerfield, 1987; Auer Jr. and Bernstein, 1997), and it has been shown that more profound lipreaders also use this information more (Pandey et al., 1986; Auer and Bernstein, 2007). Additionally, visual speech signals that convey more content information (like bilabial vs. dorsal syllables, as explained above) bias the speech percept more strongly (McGurk and MacDonald, 1976; van Wassenhove et al., 2005). However, the influence of visual information on auditory perception often depends not only on the nature and quality of the visual signal, but also on the quality of the auditory signal, since visual input is especially useful for sound identification when background noise levels are high (Sumby and Pollack, 1954; Grant et al., 2004). Thus, during audiovisual identification unimodal cues seem to be weighted based on their reliability, to create the audio-visual percept (Massaro, 1987, 1997). Additionally, the amount of weight allocated to each modality depends not only on the overall quality of the signal, but also on the reliability of the signal for the specific feature that needs to be identified. For example, spatial perception is more accurate in the visual domain, therefore spatial localization of audio-visual stimuli mostly dependents on visual signals (Driver, 1996). One of the aims of our study was to provide further support for the notion that reliable modalities are weighted more heavily (Massaro, 1997; Beauchamp et al., 2004). Specifically, we investigated whether systematic difference in the reliability of the voicing and POA features of the syllable (see above) biases which modality is weighted more heavily.

The main aim of our study was to investigate how the temporal relation between audio-visual pairs influences our percept. It is known that auditory and visual signals are only integrated when they are presented within a certain temporal window (Welch and Warren, 1986; Massaro et al., 1996; Ernst and Bülthoff, 2004), this is the so-called temporal window of integration (TWI). The TWI is for example measurable with synchrony judgments, in which temporal synchrony of audio-visual signals is only perceived if audio-visual pairs are presented within a certain range of onset asynchronies (Meredith et al., 1987; Spence and Squire, 2003). The TWI highlights that the temporal relationship of auditory and visual inputs is another important determinant for integration, in addition to information about the "what" of a stimulus. The importance of this window has been replicated many times for perceptual as well as neuronal integration (Stein and Meredith, 1993; van Atteveldt et al., 2007; van Wassenhove et al., 2007). Typical for the TWI is that the point of maximal integration occurs with visual stimuli leading (Zampini et al., 2003). This seems to relate to the temporal information visual signals provide, namely a prediction of the "when" of the auditory signal, since they naturally precede the sounds (Chandrasekaran et al., 2009; Zion Golumbic et al., 2013). However, the difference between the onset of the visual and auditory signal varies across syllables (Chandrasekaran et al., 2009) and it is not known whether these natural onset differences can cue the identity of the speech sound. It has been shown that monkey auditory cortex and superior temporal cortex are sensitive to natural audio-visual onset differences in monkey vocals (Ghazanfar et al., 2005; Chandrasekaran and Ghazanfar, 2009). In humans, it has been shown that onset differences within the auditory modality are used to identify auditory syllables (Miller, 1977; Munhall and Vatikiotis-Bateson, 1998). For example, the distinction between a voiced or unvoiced syllable in the auditory signal is solely based on onset differences of specific frequency bands. However, it is not known whether audio-visual onset information is used to identify speech sounds. We hypothesize that inherent onset differences between auditory and visual articulatory cues can be used to identify spoken syllables. Specifically, we hypothesize that coronal (e.g., /da/) and dorsal (e.g., /ga/) stimuli (pronounced with the front or body of the tongue, respectively) might have audio-visual onset difference, in which dorsal stimuli produce longer onset differences due to a longer distance from the POA to the external, audible sound.

Traditionally, only a single dimension in the auditory or visual signal is altered to investigate the influence of visual cues. However, more and more studies are showing interactions between different crossmodal cues. For example, Vatakis and Spence (2007) found that if the gender of a speaker is incongruent for auditory and visual speech, less temporal discrepancy is allowed for the stimuli to be perceived as synchronous. Stimuli in the McGurk effect (McGurk and MacDonald, 1976), in which an auditory [ba], presented with an incongruent visual /ga/ is perceived as a /da/, are also perceived as synchronous for a narrower temporal window, compared to congruent audio-visual syllables (van Wassenhove et al., 2007). Furthermore, in recent work we showed that auditory detection thresholds are lower if temporal predictive cues are available in both the auditory and visual domain (ten Oever et al., submitted). In addition, interactions between semantic relatedness and spatial processing have been reported (Driver, 1996; Parise and Spence, 2009; Bien et al., 2012), as well as interactions between temporal and spatial factors (Stevenson et al., 2012). However, it is still unknown how interactions between auditory and visual content as well as temporal cues influence speech identification.

In sum, for stop consonants, auditory cues provide content information with regard to voicing, whereas visual cues provide content information with regard to POA (with varying reliability, e.g., for bilabial vs. dorsal/coronal). Therefore, we were able to make use of these properties in order to investigate whether incongruent pairs of stimuli are identified depending on the modality

that has the most reliable information for the specific features; POA and voicing. Additionally, we used different SOAs to investigate the temporal profile of this effect. Specifically, we were interested in the temporal window in which visual information influences the auditory percept, and whether ambiguity in the identity of auditory syllables can be resolved using differences in natural audio-visual onsets in speech.

MATERIALS AND METHODS

PARTICIPANTS

Eight healthy native Dutch volunteers (3 male, mean age 20.9, SD 2.6) participated in the study. All participants reported to have normal hearing and normal or corrected to normal vision. Participants were unaware of the goal of the study before they completed the experiment. Informed consent was given before participating. Ethical approval was given by the Ethical Committee of the Faculty of Psychology at the University of Maastricht. Participants received €40 or student participation credits in compensation for their time.

STIMULUS MATERIAL

Six Dutch syllables, pronounced by a native Dutch female speaker, were used as auditory and visual stimuli (/pa/, /ba/, /ta/, /da/, /ka/, /ga/). For variability, we recorded three different versions of every syllable. Sounds were digitized at 44.1 kHz, with 16-bit amplitude resolution and were equalized for maximal intensity. Videos had a digitization rate of 30 frames per second and were 300 × 300 pixels. We used a method similar to method used in van Wassenhove et al. (2005) to create the videos. Videos lasted 2367 ms, including a fade in of a still face (8 frames), the still face (5 frames), the mouth movements (52 frames), and a fade out of a still face (5 frames). MATLAB (Mathworks) scripts were used to create these videos. Additionally, for every stimulus there was a still face video with the fade out and fade-in frames. First, we tested three participants with SOAs between auditory and visual stimuli ranging from VA (visual lead) 300 ms up to AV (auditory lead) 300 in steps of 30 ms, since this range covers the TWI for syllables used before (see e.g., van Wassenhove et al., 2007; Vatakis and Spence, 2007). However, for the extreme VA and AV SOAs participants still seemed to use the visual information to determine their responses, therefore we chose to widen the SOA range (ranging from VA 540 to AV 540 ms in steps of 60 ms for the other participants). To align the incongruent auditory stimuli with the videos, the maximal intensity of the incongruent auditory stimulus was aligned with the congruent auditory stimulus.

PROCEDURE

Each participant was tested in two separate experimental sessions, both lasting 2 h. In the first session a staircase, a unimodal visual experiment, and the first part of the audio-visual experiment was conducted. The second session consisted of the remainder of the audio-visual experiment.

The staircase procedure consisted of a six-alternatives forced choice procedure in which participants were asked to identify the six different syllables without presentation of the videos. Syllables were randomly presented over a background of white noise. Depending on the accuracy of the response, the intensity of the white noise was increased or decreased for the next trial. A two-up, one-down procedure (Levitt, 1971) with a total of 20 reversals was employed, which equals approximately 70% identification threshold. The individually obtained white noise intensity was used in the following experiments as background noise for the individual participants.

In the unimodal visual experiment participants were requested to recognize the identity of the syllable based on the videos only. White noise was presented as background noise. First, a fixation cross was presented for 800 ms, followed by a syllable video. Finally, a question mark was presented with the six possible response options to which participants were requested to respond. After participants responded there was a 200-ms break before the next trial started. In total, 360 stimuli were presented, 60 per syllable in 4 separate blocks.

The audio-visual experiment had a similar trial configuration to the unimodal visual experiment, but consisted of the presentation of audio-visual pairs. Only two visual stimuli were used here; /pa/ and /ga/. These specific syllables were selected because they differ from each other in terms of POA: /pa/ is a bilabial syllable, pronounced in the front of the mouth, whereas /ga/ is dorsal syllable, pronounced in the back of the mouth. Furthermore, it has been shown that identifying /pa/ is much easier than /ga/ (Wiener and Miller, 1946; McGurk and MacDonald, 1976; van Wassenhove et al., 2005), thus serving our aim to manipulate the amount of information provided by the visual stimulus. Participants were instructed to identify the auditory stimulus only (again choosing between the six possible response options), while ignoring the identity of the visual stimulus.

In total, 30 blocks were presented, distributed across the two sessions for all participants. Furthermore, per SOA there were 10 stimuli for every audio-visual combination for the five participants who saw the full range of SOAs, and 11 stimuli per SOA for the other three participants. Blocks lasted approximately 7 min each. Additionally, there were catch trials in which a visual or auditory unimodal stimulus (20 stimuli for each) was presented. During the auditory unimodal presentation randomly one of the still visual faces, which were also used during the fade in of the moving faces, was presented. During the visual unimodal presentation white noise was presented at the same intensity as the audio-visual trials and participants had to indicate the identity of the visual stimulus. This ensured that participants were actually looking at the screen.

Participants were seated approximately 57 cm from the screen and were instructed to look at the fixation cross at all times if presented. Presentation software (Neurobehavioral Systems, Inc., Albany, NY, USA) was used for stimulus presentation. Visual stimuli were presented on a gray background (RGB: 100, 100, 100). After each block participants were encouraged to take a break and it was ensured that participants never engaged continuously in the task for more than half an hour.

DATA ANALYSIS

With regard to the unimodal stimuli, we aimed to replicate previous findings stating that voicing is discriminated better in the auditory modality, whereas POA is discriminated better in the visual modality (Wiener and Miller, 1946; McGurk and MacDonald, 1976; Summerfield, 1987). For the analysis

concerning voicing, the percentage of voiced responses was calculated per voicing category. Thereafter, we averaged the response proportions and performed an arcsine-square-root transformation to overcome non-normality caused by the restricted range of the proportion data (however in the figures proportions are kept for illustration purposes, since they are more intuitive). The calculated transformed response proportions per category were used as dependent variables in two repeated measurements ANOVAs, for the visual as well as for the auditory modality. For the visual unimodal analyses, the data from the unimodal visual experiment was used (although the data from the visual catch trials in the AV experiment gave comparable results), whereas for the auditory analyses the catch trials in the audio-visual experiment were analyzed. To investigate whether participants could identify the voicing of the stimulus the factors Voicing of the stimulus (voiced vs. unvoiced stimuli) and Voicing of the response were used. A similar analysis was performed to investigate whether POA could be identified in the auditory and visual modality. Here, the percentage of POA responses per POA category were calculated, arcsine-squared-root transformed, and the factors POA of the stimulus (bilabial, coronal, or dorsal) and POA of the response were used in two repeated measurements ANOVAs for the visual and auditory modality. For significant interactions simple effect analyses per stimulus category were performed. If not otherwise reported, all multiple comparisons were Bonferroni corrected and effects of repeated measures were corrected for sphericity issues by Greenhouse–Geisser correcting the degrees of freedom.

For the Audio-visual analyses, we first performed the same analyses as for the unimodal stimuli, collapsed over the SOAs, separately for visual /pa/ and /ga/. Thereafter, linear mixed models were used to investigate the SOA effects. This approach was chosen to accommodate for the missing data which arose because three participants were only presented with SOAs between VA 300 and AV 300 ms instead of VA 540–AV 540 ms. Per visual stimulus and per voicing level a mixed model was run with the factors Stimulus POA, Response (only responses that were on average per VC category above chance level were used for further analyses) and SOA. This factor was created by binning the differently used SOAs in nine bins with center points: VA 50, 125, 275, and 475, 0 and AV 50, 125, 275, and 475. These bins were chosen to include all the SOAs used. Additionally, a random intercept was added to account for the individual variations in the baseline.

We hypothesized differential effects as a result of natural differences in onset asynchronies of mouth movements and congruent speech sounds, for example between dorsal (earlier movements) and coronal syllables (later movements). In order to investigate this hypothesis, we calculated the velocity of the mouth movements as follows. For each visual stimulus we zoomed in on the area around the mouth (see **Figure 1**). Then, the mean of the absolute differences of the three RGB values per pixel for adjacent frames was calculated. Thereafter, to quantify the movement from one frame to the other, the variance of the mean absolute RGB differences over the pixels was calculated and this was repeated for all the frames. This resulted in a velocity envelope of the mouth movement (i.e., comparable to the derivative of the mouth movement—it indicates *changes* in the movement) in which a clear opening and closing of the mouth becomes visible (see **Figure 1**). The result of this method is similar to the methods used by Chandrasekaran et al. (2009), such that the point of maximum velocity coincides with a half open mouth and the minimum velocity coincides with a fully open mouth. To quantify the onset differences between the auditory and visual signals, the time point of maximal amplitude of the auditory signal was subtracted from the time point of maximal velocity of the visual signal. These values were later used in a linear mixed model (see Results for details).

RESULTS

UNIMODAL EFFECTS

We replicated previous results showing that voicing is most optimally discriminated in auditory syllables and POA most optimally in visual syllables (see **Figure 2**; **Tables 1** and **3**). **Table 1** indicates that the response POA interacts with the stimulus POA only for the visual stimuli, which means that for a stimulus with a specific POA the POA categories have different response proportions during the visual experiment. Simple effects show that especially bilabial stimuli were identified correctly during the visual experiment (as indicated by significantly higher bilabial than dorsal and coronal responses). Dorsal and coronal visual stimuli were more often confused with each other. However, for the unimodal auditory stimuli, the interaction between response and stimulus

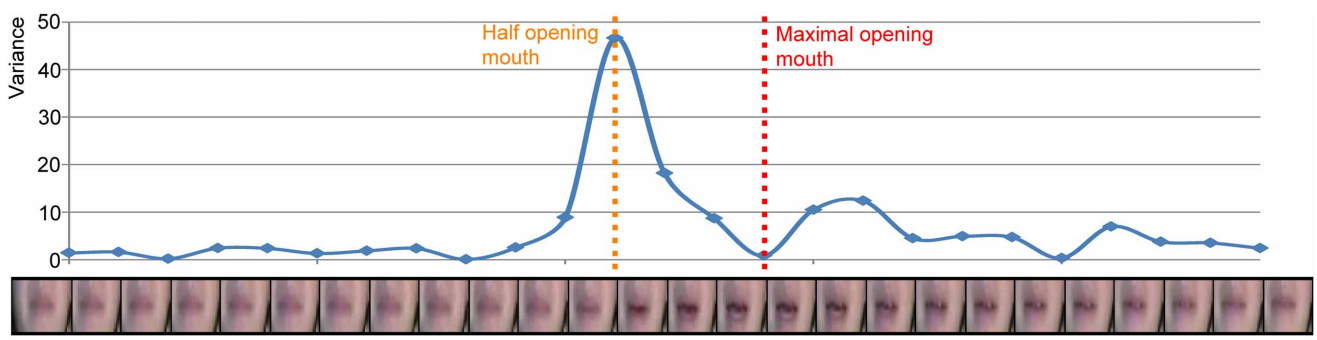


FIGURE 1 | Example of the envelope of the velocity of the mouth movement of visual /pa/. Each dot represents the variance over all pixels of the mean RGB difference for two adjacent frames (the frame left and right of the dot). The orange dotted line represents the half opening of the mouth and the red dotted line represents the maximal opening of the mouth.

POA did not reach significance, indicating that participants were not able to dissociate the POA of the auditory stimuli. **Table 3** (top rows) shows significant simple effects of the voicing of the response per stimulus level for the auditory, but not the visual modality. This means that in the auditory modality, voicing was primarily categorized correctly.

MULTIMODAL EFFECTS COLLAPSED OVER SOAs

During the audio-visual experiment, the voicing of the stimuli was identified correctly most of the time (as indicated by significant simple effects for the voicing analyses; see **Figure 3**; **Table 3**), and resembles the results from the unimodal auditory analyses. The results for the POA, when visual /pa/ was presented, resulted in high response proportions (more than 0.8) for bilabial stimuli (see **Table 2**), paralleling visual unimodal results. The POA response × stimulus interaction effect indicates that bilabial responses are specifically reported when the auditory stimuli is also bilabial, but in the simple effects the comparisons did not show significant differences (**Table 2**, row 3). Similarly, the response distributions for dorsal stimuli in the unimodal visual experiment and the visual /ga/ during the audio-visual experiment seem to resemble each other, that is, in the audio-visual experiment participants also confused the coronal and dorsal POA.

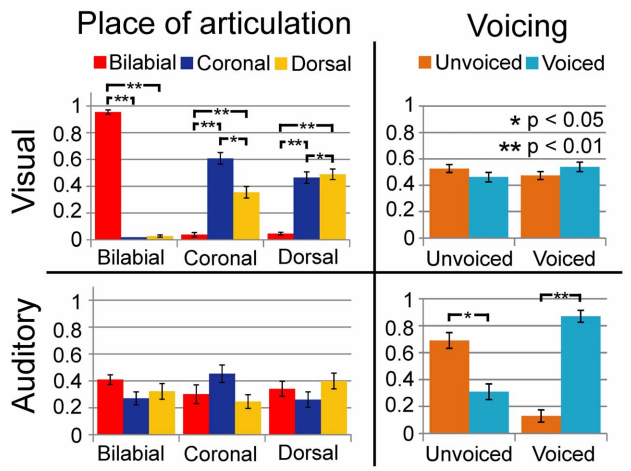


FIGURE 2 | Results of unimodal analyses for auditory and visual signals separately. Horizontal axis represents the category of the stimulus and vertical axis represents the response proportions of the respective categories. Dashed lines indicate chance level performance. As shown, vision can dissociate place of articulation (POA) and audition can dissociate voicing (VC).

The latter analysis shows that adding a visual stimulus changes the auditory percept for the different POA categories, such that with incongruent audio-visual POA, the correct POA response choice (i.e., the POA of the auditory stimulus) is nearly absent in the chosen responses. For example, although a dorsal auditory stimulus is presented (e.g., /ka/), if concurrently visual /pa/ is presented, the response options with dorsal POAs are only chosen approximately 10% of the times (see **Figures 3** and **4**). Therefore, we decided that, for the analyses including the temporal factors, we would only use the response options that were given more than chance level per stimulus voicing and POA (POA: 0.33, voicing: 0.5). Mainly, because we were interested in the temporal pattern of the identification and a very low response rate could result in floor effects, biasing the statistical analyses. Thus for visual /pa/, auditory unvoiced we only used response /pa/ (see **Figure 3**; stimulus unvoiced and POA bilabial) and for visual /pa/, auditory-voiced we only used response /ba/ (stimulus voiced and POA bilabial). For visual /ga/, auditory-unvoiced response options /ta/ and /ka/ were used (stimulus unvoiced and POA coronal and dorsal respectively) and for visual /ga/, auditory-voiced response options /da/ and /ga/ were used (stimulus voiced and POA coronal and dorsal respectively).

TEMPORAL EFFECTS DURING VISUAL /pa/

Overall effects of SOA difference are shown in **Figure 4**. The mixed model analyses for visual /pa/, auditory unvoiced showed an main effect for POA and SOA [**Figure 5A**; $F(2, 180) = 34.04$, $p < 0.001$ and $F(8, 180) = 10.88$, $p < 0.001$, respectively]. Bilabial responses were reported significantly more than coronal and

Table 1 | Results for the POA analyses of the unimodal stimuli.

(A)		POA interaction	Simple effects per stimulus level								
			Stimulus bilabial (B)			Stimulus coronal (C)			Stimulus dorsal (D)		
			B vs. C	B vs. D	C vs. D	B vs. C	B vs. D	C vs. D	B vs. C	B vs. D	C vs. D
Auditory	F/t	2.34	–	–	–	–	–	–	–	–	–
	P	0.12									
Visual	*F/t*	178.4	23.2	26.8	−0.92	−9.89	−8.24	2.70	−9.6	−13.1	−0.16
	P	0.00**	0.00**	0.00**	1.00	0.00**	0.00**	0.09	0.00**	0.00**	1.00

*The second column shows the interaction between stimulus and response place of articulation (POA interaction), and the other three columns show for stimuli with the different POAs the pairwise comparisons of the response proportions between the different POAs responses (B, bilabial; C, coronal; and D, dorsal). Auditory and visual rows indicate the results from the auditory only trials during the audio-visual experiment and the separate unimodal visual experiment, respectively. Results for post hoc analyses are only shown if ANOVA tests are significant. **indicates p-values below 0.01.*

Table 2 | Results for the POA analyses of the multimodal stimuli.

(B)		POA interaction	Simple effect for congruent response			POA Response; main effect	Pairwise comparisons of response level		
			B vs. C	B vs. D	C vs. D		B vs. C	B vs. D	C vs. D
AV, visual /pa/	*F/t*	6.30	2.41	2.23	−1.89	92.2	8.33	10.6	1.15
		0.02*	0.14	0.19	0 29	0.00**	0.00**	0.00**	1.00
AV, visual /ga/	*F/t*	3.43	–	–	–	39.78	−4.80	−7.94	0 03
	p	0.07				0.00**	0.01**	0 00**	1.00

*The second column is similar as in **Table 1**. The third column shows the simple effect for the visual congruent response option (for visual /pa/ the bilabial response), comparing whether for specific stimuli the congruent visual POA option has a higher proportion. The fourth column shows the main effect of the response of the POA. The last column shows the pairwise comparisons whether overall, one POA response is given more often than another (B, bilabial; C, coronal; and D, dorsal). Results for post hoc analyses are only shown if ANOVA tests are significant. * and ** indicate p-values below 0.05 and 0.01, respectively.*

Table 3 | Results for voicing for both unimodal and multimodal stimuli.

(C)		Voicing interaction	Response simple effects per stimulus level: voiced vs. unvoiced	
			Stimulus voiced	Stimulus unvoiced
Auditory	*F/t*	43.8	8.19	−2.83
	p	0.00**	0.00**	0.03*
Visual	*F/t*	18.5	1.66	−0.13
	p	0.00*	0 14	0.90
AV, visual /pa/	*F/t*	112	8.71	−6.82
	p	0.00**	0.00**	0.00**
AV, visual /ga/	*F/t*	87.2	11.42	−3.94
	p	0 00**	0.00**	0.01**

*The second column is the interaction of stimulus voicing with response voicing (voicing interaction). The third and fourth columns are the simple effect analyses of the voicing of the response per stimulus level. Results for post hoc analyses are only shown if ANOVA tests are significant. * and ** indicate p-values below 0.05 and 0.01, respectively.*

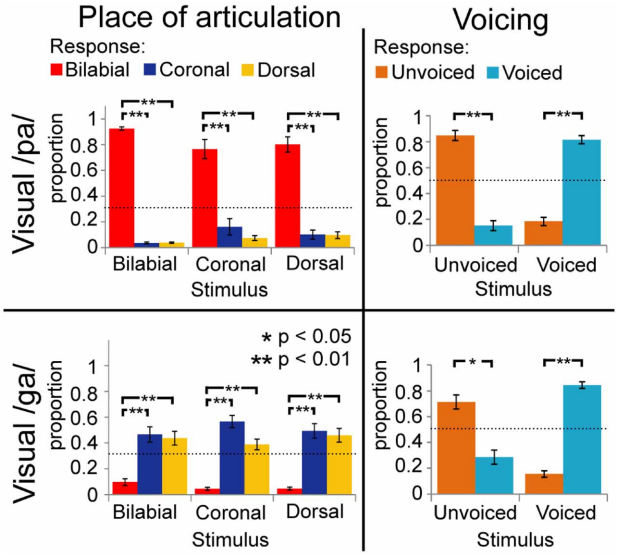


FIGURE 3 | Results of multimodal analyses for visual /pa/ and /ga/ separately collapsed over stimulus onset asynchronies (SOAs). Horizontal axis represents the category of the stimulus and vertical axis represents the response proportions of the respective categories. Dashed lines indicate chance level of responding. Voicing (VC) is dissociable, but place of articulation (POA) responses depended on the unimodal visual response in **Figure 2**.

dorsal responses [$t(180) = 7.60$, $p < 0.004$ and $t(180) = 6.59$, $p < 0.001$, respectively]. The main effect of SOA indicated that compared to an SOA of zero, for AV 475 and AV 275 lower /pa/ response proportion were given [$t(180) = -4.60$, $p < 0.001$ and $t(180) = -4.583$, $p < 0.001$, respectively]. Thus, the proportion /pa/ responses were the least for incongruent bilabial presentation, and when auditory stimuli were leading more than 125 ms. Visual /pa/, auditory-voiced stimuli resulted in similar results: an main effect for POA and SOA [**Figure 5B**; $F(2, 180) = 13.59, p < 0.001$ and $F(8, 180) = 4.83$, $p < 0.001$, respectively]. Bilabial response proportions were higher than coronal and dorsal response proportions [$t(180) = -4.49, p < 0.001$ and $t(180) = -4.54, p < 0.001$, respectively]. Here, for a smaller window /ba/ responses were given compared to visual /pa/–unvoiced /pa/ responses, that is, the SOAs of AV 475, AV 275, and VA 475 were significantly different from an SOA of zero [AV 475: $t(180) = -4.027$, $p < 0.001$; AV 275: $t(180) = -3.639$, $p = 0.003$; and VA 475: $t(180) = -3.584$, $p = 0.004$].

TEMPORAL EFFECTS DURING VISUAL /ga/

The multilevel analyses for the visual /ga/ unvoiced showed an interaction effect between response and SOA [$F(8, 371) = 4.540$, $p < 0.001$]. Results from the simple effects analyses in which the /ta/ and /ka/ responses per SOA level were compared indicated that for SOA VA 275 /ka/ was indicated more and for SOA AV 50, 125, and 475 /ta/ was indicated more [uncorrected values: $-275 = -2.813$, $p = 0.008$; 50: $t(24) = 2.088$, $p = 0.041$; 125: $t(24) = 2.394$, $p = 0.022$; 475: $t(24) = 2.650$, $p = 0.014$], but these effects did not survive correction for multiple comparisons. The interaction effect however, seems to be caused by more answered /ka/ with negative SOAs, and more answered /ta/ with positive SOAs (see **Figure 6A**).

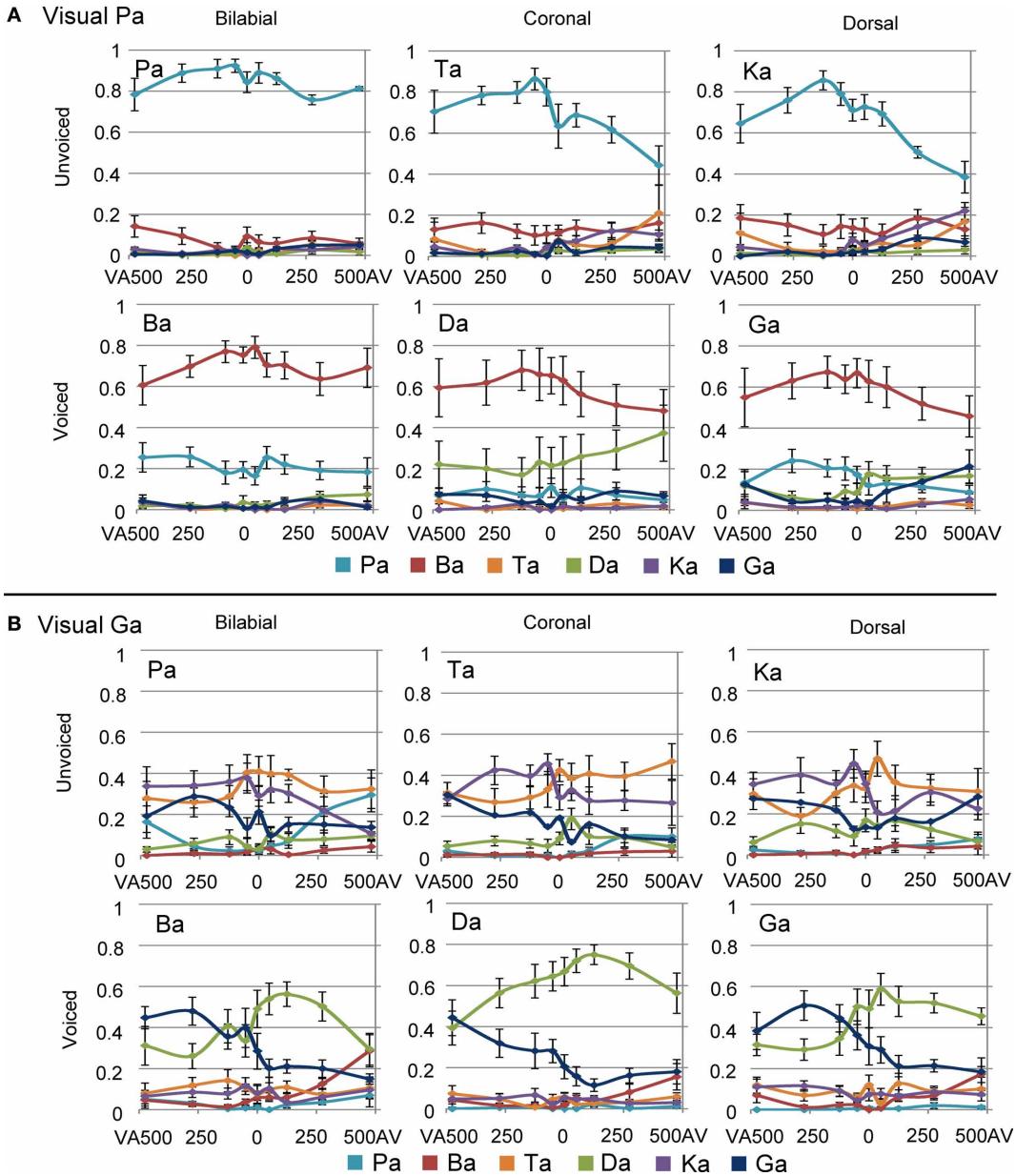


FIGURE 4 | Overall results of the multimodal experiment for visual /pa/ (A) and visual /ga/ (B), combined with the six auditory stimuli and all stimulus onset asynchronies (SOAs). Negative SOAs indicate that the visual stimulus was shifted to an earlier point in time compared to the auditory stimulus.

For the visual /ga/, auditory-voiced the multilevel analyses also showed an interaction of response and SOA [see **Figure 6B**; $F(8, 367) = 11.996$, $p < 0.001$]. Additionally, it showed an interaction between stimulus POA and response [$F(8, 367) = 26.480$, $p < 0.001$]. One explanation for this last effect could be that our [da] stimulus was better identifiable unimodally than the other auditory stimuli (see **Figure 4**), such that for stimulus POA coronal a higher proportion /da/ responses were given (since this was the right answer). This was similar during visual /pa/, auditory [da], which also showed a higher proportion /da/ compared to the correct responses during other incongruent combinations (**Figure 4A**). For the response × SOA interaction we performed simple effects analyses per SOA level. For all AV SOAs and SOA 0 /da/ was reported significantly more than /ga/ [475: $t(24) = 4.667$, $p < 0.001$; 275: $t(24) = 7.624$, $p < 0.001$; 125: $t(24) = 9.089$, $p < 0.001$; 50: $t(24) = 6.615$, $p < 0.0001$; 0: $t(24) = 3.922$, $p = 0.004$].

"CROSSING" IDENTIFICATION FOR VISUAL /ga/

Around the zero point, we observed a quick incline or decline in the response choice of participants for visual /ga/ (see **Figures 4B** and **6**), such that participants chose with positive SOAs more

often coronal responses (/da/ or /ta/) and with negative SOAs more often dorsal responses (/ga/ or /ka/). The decline seems to be less strong for visual /ga/, auditory [da]. This is probably related to the better unimodal auditory identification of auditory [da]. However, also here the incline for /ga/ responses and decline for /da/ responses around zero is observable. The "crossing" could relate to inherent differences in onsets between visual and auditory signals for coronal and dorsal stimuli. Indeed, a 2 × 3 ANOVA with factors POA and VC comparing onset differences between the maximal amplitude for visual velocity and auditory signal showed an effect of POAs [see **Figure 6C**; $F(1, 12) = 8.600$, $p = 0.005$]. Pairwise comparisons showed that dorsal stimuli had significantly bigger AV onset differences than coronal or bilabial stimuli [dorsal-coronal: $t(5) = 2.757$, $p = 0.012$; dorsal-bilabial: $t(5) = 1.941$, $p = 0.033$; bilabial-coronal: $t(5) = 0.466$, $p = 1.000$]. In our stimulus set we did not find a significant difference between voiced and unvoiced stimuli [$F(1, 12) = 0.800$, $p = 0.389$], so we collapsed this for further analyses and figures.

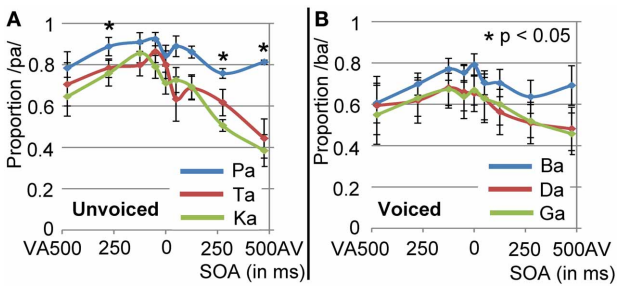


FIGURE 5 | Results for visual /pa/ presentation for unvoiced stimuli (A) and voiced (B) stimuli. Only response proportions are shown for the response options that were given above chance level. These response options were /pa/ and /ba/, for unvoiced and voiced stimuli respectively.

To model whether these inherent differences in onset asynchronies could explain the observed crossing, a new mixed model analysis was conducted. Therefore, we changed the factor SOA into a quantitative factor as described in **Figure 6D**. The logic of the model is as follows: since both unimodal stimuli alone cannot conclusively define the identity of the stimulus (auditory unimodal can differentiate voicing, but visual unimodal can only exclude bilabial), two options are left. Our perceptual system might resolve this issue by using another cue, namely time differences between audio-visual syllable pairs. In our stimulus set, a SOA of zero is equal to the onset asynchronies of dorsal stimuli, because we aligned the stimuli based on the maximal amplitude of auditory [ga] (see **Figures 6C,D**). The difference between dorsal and coronal onsets is on average 80 ms (average audio-visual asynchrony for dorsal is 135 ms and for coronal 55 ms). Therefore, the SOA for coronal stimuli in our stimulus set would be around +80 ms. With SOAs bigger than 80 ms the onset asynchronies match closer to coronal than to dorsal asynchronies. The opposite is true for audio-visual pairs with a long (experimental) visual lead: the onset asynchronies are close to dorsal asynchronies. In between these natural lags there is an ambiguity with regard to the identity of

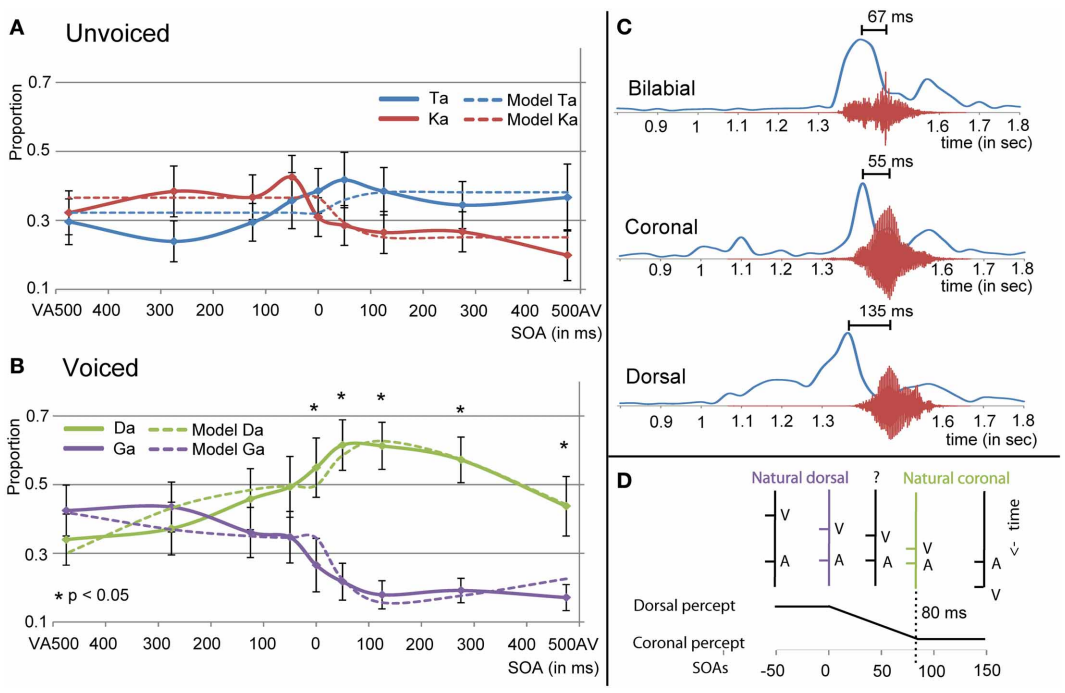


FIGURE 6 | Results for visual /ga/ presentation for unvoiced (A) and voiced (B) auditory stimuli. (C) Shows the onset differences in visual velocity and auditory amplitude for the different place of articulations (POAs). **(D)** Shows the predictor for the mixed model analyses using the natural dorsal and coronal onset asynchronies. The fit of the model together with the other significant predictors in the mixed model analyses is represented in **(A,B)** as dashed lines.

the stimulus. This factor therefore specifically tests our hypothesis that dependent on the audio-visual onset difference, participants would be biased in choosing the dorsal or coronal option, which provides new insight in the mechanism of how the percept is formed in case of ambiguous inputs. Additionally, we added a second order polynomial to the analyses to account for the downslope at the extremes.

The results of this mixed model showed an interaction between response and the created factor in both the unvoiced and voiced analyses [**Figure 6B**; $F(1, 385) = 22.446$, $p < 0.001$ and $F(1, 379) = 58.166$, $p < 0.001$, respectively], indicating that indeed modeling the natural lag in audio-visual syllables explains the difference in the response choice for the different SOA. In both voicing levels dorsal responses had positive and coronal responses negative values for the parameter estimate (Unvoiced: parameter estimate −0.1410 and 0.0689 for /ta/ and /ka/ respectively and Voiced: parameter estimate −0.2212 and 0.1674 for /da/ and /ga/ respectively), verifying the hypothesized pattern of the effect in which negative SOAs should result in a dorsal percept. As in the previous analyses, POA showed an interaction with response for the visual /ga/ stimulus [$F(2, 379) = 26.731$, $p < 0.001$]. The second order factor was only of significance in the analyses with the voiced stimuli and showed an interaction with response [$F(1, 379) = 22.279, p < 0.001$], such that the parameter estimate was more negative for the /ga/ response.

DISCUSSION

The current study investigated the influence of content and temporal cues on the identification of audio-visual syllables. We hypothesized that visual input influences the percept only within a constrained temporal window. Furthermore, we predicted that the more reliable unimodal content cues determine the percept more strongly. Finally, we hypothesized that information about natural audio-visual onset differences can be used to identify syllables. We revealed that during audio-visual speech perception visual input determines the POA and auditory input determines the voicing. Moreover, we confirmed the prediction of a wide window in which visual information influences auditory perception that was wider for congruent stimulus pairs. Interestingly, within this window, the syllable percept was not consistent, but differed depending on the specific SOA. This was particularly pronounced when the POA could not be reliably identified (i.e., between dorsal and coronal stimuli). We explained this temporal response profile using information about natural onset differences between the auditory and visual speech signals, which are indeed different for the dorsal and coronal syllables.

MULTIPLE UNIMODAL CUES FOR AUDIO-VISUAL SPEECH IDENTIFICATION

Our data suggest that participants used the visual signal to identify the POA and the auditory signal to identify voicing during audio-visual presentation. We suggest that it is the reliability of the cue for the specific features of the syllable that determined the percept, since it has been shown before that the reliability of a cue can determine the percept (Massaro, 1997; Andersen et al., 2004). This is also in line with our replication of the results that unimodally, visual stimuli are best dissociable by using POA and auditory stimuli are best dissociable by using voicing (Wiener and Miller, 1946; Summerfield, 1987; van Wassenhove et al., 2005). It appears that irrespective of the task, which was to identify the auditory stimulus, visual input influences perception. Therefore, it seems that audio-visual speech is automatically integrated, since participants were not able to perform the task using only auditory cues as instructed. Integration in our study is shown by different identification responses for auditory and audio-visual presentation of the same spoken syllables. This perceptual effect is similar to the McGurk effect, in which identification of an auditory syllable is involuntarily influenced by an incongruent visual input (Soto-Faraco et al., 2004; Gentilucci and Cattaneo, 2005). This indicates that during audio-visual speech perception, an integrated percept is created that uses the information of the visual as well as the auditory domain. In the current setting, since the auditory signal is non-optimal, this leads to a considerable bias in favor of the visual POA, for which the visual input is most reliable and thus dominant. In the McGurk effect, both signals are equally salient, resulting in a fused percept. So, when a unimodal signal is dominant during audio-visual integration, this predisposes perception.

CONTENT PREDICTIONS IN AUDIO-VISUAL SPEECH

In the current study we manipulated the predictability of the visual signal by using one visual syllable in which the POA can reliably be determined (/pa/) and another syllable in which the POA estimate is less reliable (/ga/). Previous research has shown that the information present in the visual signal is used to determine our percept, for example, van Wassenhove et al. (2005) showed facilitation of congruent speech dependent the amount of content information in the visual stimuli. Consistent with our results, van Wassenhove and colleagues showed that, /pa/ stimuli which convey more content information about POA, influenced electro-encephalographic recordings more than a less informative syllable /ka/. In their study, an analyses-by-synthesis framework was proposed in which the auditory signal is evaluated, based on the predictive strength the visual signal has for the content of the auditory signal. This predictive strength should determine whether there is a McGurk effect (van Wassenhove et al., 2005) and should also correlate with prediction error when an incongruent auditory stimulus is presented (Arnal et al., 2011). In a study using congruent audio-visual speech with auditory speech in white noise, Pandey et al. (1986) showed that more proficient lip readers can still detect the auditory signal at higher noise levels, indicating that the predictive strength or the amount of information conveyed by the visual signal, influences the amount of benefit during auditory perception. Here, we also show that more predictable visual bilabial stimuli bias the percept more strongly, because visual /pa/ shaped the percept more profoundly than visual /ga/. This is in line with results from Vatakis et al. (2012) who found that the point of perceived synchrony needed more visual lead for stimuli pronounced more in the back of the mouth compared to bilabial stimuli. They argue that for more salient visual stimuli (i.e., bilabial stimuli) a smaller visual lead is required to reach synchrony perception. In our study, this is reflected in the amount of bias of the visual signal for the POA response choice. Since the auditory signal had a low signal to noise ratio, the visual signal biases the percept of POA

completely, such that unimodal and audio-visual POA response proportions were the same.

INTERPLAY BETWEEN TWO DISTINCT TEMPORAL CUES IN AUDIO-VISUAL SPEECH PERCEPTION

It is well-known that temporal cues are informative for audiovisual speech identification (Munhall and Vatikiotis-Bateson, 2004; Zion Golumbic et al., 2012). Firstly, auditory and visual speech seems to temporally co-vary (Campbell, 2008). Especially in theta frequencies around 2–7 Hz, lip movement and the auditory envelope seem to correlate (Müller and MacLeod, 1982; Chandrasekaran et al., 2009; Luo et al., 2010). This feature has been considered a main source of binding and of the parsing of information (Poeppel, 2003; Campbell, 2008; Ghazanfar et al., 2013) and removing this frequency reduces auditory intelligibility (Vitkovitch and Barber, 1994; Ghitza, 2012). Secondly, visual signals generally precede auditory signals, providing temporal predictability of the arrival of the auditory signal (Schroeder et al., 2008). Finally, audio-visual speech perception has generally been shown to have a broad integration window (Dixon and Spitz, 1980; Grant and Greenberg, 2001), which has led to the conclusion that audio-visual speech perception has loose temporal associations (Munhall and Vatikiotis-Bateson, 2004). Our results also indicate that visual input influences the auditory percept for a wide range of SOAs. For example, we show that with auditory [ba] and visual /ga/, the visual signal influences the percept for a time window in which the visual signal is shifted 500 ms earlier in time, relative to the auditory signal, up to when the visual signal was shifted 300 ms later in time, relative to the auditory signal (SOAs ranging from VA 500 up to AV 300 ms). Only at the most positive SOA (AV 500) is visual information not used and the correct answer [ba] is present in the given responses.

Although we find integration during a wide window, the results do not support a very loose temporal association, since we also found evidence for the use of natural temporal audio-visual onset differences in identifying the syllable. However, this information was only used when unimodal cues did not provide enough information. Therefore, we propose the following mechanism for the interplay of articulatory cues (POA and voicing), temporal integration cues, and temporal onset cues (see **Figure 7**): first, the visual and auditory components of a syllable activate syllable representations based on their "preferred" cue and reliability. However, these activations have some decay, such that at some point in time after the visual stimulus was presented, visual information does not influence the percept anymore (the TWI). Within this window more reliable cues will cause more activation of specific representations (i.e., visual cues will activate representations of syllables with corresponding POAs and auditory cues will activate representations of syllables with corresponding voicing). In a winner-takes-all framework, which is the case in an identification task, only one representation can win and that will be the representation with the strongest input. However, in addition to the visual and auditory articulatory cues, the activation of syllable representation is also based on the encoded natural onset differences. That is, for dorsal stimuli (e.g., /ga/), maximal activation will occur later than for coronal stimuli (e.g., /da/). When an ambiguous auditory

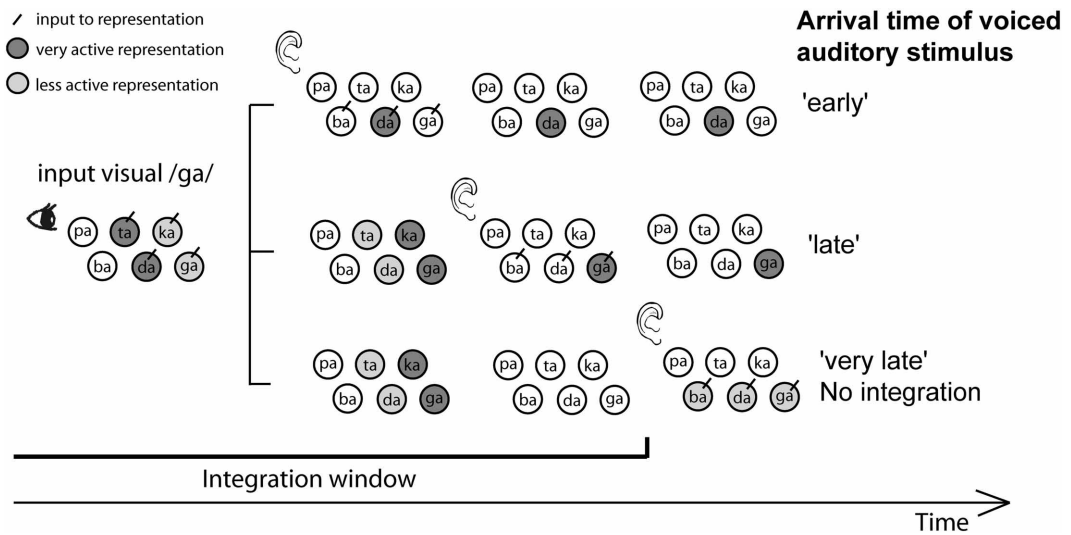


FIGURE 7 | Proposed mechanism explaining the interplay between place of articulation (POA), voicing, temporal integration, and temporal onset cues. The figure shows what happens during the presentation of visual /ga/ and an ambiguous voiced stimulus. For the visual syllable /ga/, POA cues present in the visual signal activate (indicated by darker circles) coronal (/ta/ and /da/) and dorsal (/ka/ and /ga/) representations in a time-dependent manner: the activation decays over time (indicating the TWI), and depending on the natural audio-visual onset differences, maximal activation occurs at different time points for the two POAs (later for dorsal than for coronal). Therefore, the time when auditory information activates representations of syllables (represented along the vertical axis) is important for winning the decision making process. When auditory syllables arrive early, and therefore resemble more closely natural audio-visual onset differences for /da/, /da/ is more active than /ga/, and has the highest chance to win the decision making process. In this example, the visual cues can not distinguish between the different coronal and dorsal possibilities, and the auditory cues cannot distinguish the POA at all, so the arrival of the auditory information (early vs. late) facilitates this decision; early onset will activate the coronal /da/ and late onset will activate the dorsal /ga/ syllable.

stimulus arrives, it will activate multiple representations (the three voiced representations in the figure). The representation that is most active at that point in time, depending on the audio-visual onset difference, will win the competition. In the figure, visual /ga/ input cannot dissociate the coronal (/da/ and /ta/) from the dorsal (/ga/ and /ka/) POA, and auditory information cannot dissociate the POA at all. Therefore, if the auditory stimulus arrives early (resembling natural coronal audio-visual onset differences), the most active representation will win the competition, in this example /da/. For later presentation, /ga/ will be more activated, and when the decay is completed there is no bias from the visual cue (since no representations are active), and one of the three voiced stimuli has to be chosen. This way, audio-visual onset differences only influence identification when ambiguous auditory stimuli are presented within the TWI, and only if the visual POA cues are not decisive.

TEMPORAL WINDOW OF INTEGRATION IS INFLUENCED BY AUDIO-VISUAL CONGRUENCY

The TWI is generally measured by evaluating whether participants can indicate if audio-visual events are presented simultaneously or not (Vroomen and Keetels, 2010), assuming that when participants can reliably dissociate the two, the audio-visual event is perceived as two separate events and not bound together. However, little research has been done to assess whether audio-visual SOA differences also influence unimodal perception, which was one of the aims of the current study. Applying the same logic as that used for simultaneity judgments, events that are bound should influence unimodal perception more than when they are perceived separately. We here show that especially during congruent audio-visual voicing (visual /pa/, auditory unvoiced), the response proportions of /pa/ are higher (**Figure 5**). Also, visual influence seems to have a wider TWI for the congruent pairing of visual /pa/ with auditory /pa/, as the visually determined /pa/ response proportion appears higher for a wider temporal window (although the statistical test did not show this). One explanation for these congruency effects is the "unity assumption" stating that when two stimuli naturally belong together they are bound more strongly and therefore are more difficult to dissociate over a wider temporal window (Welch and Warren, 1980). However, it could be that with extreme SOAs, visual information is not used and participants rely only on the auditory signal, that is, in the case of congruent audio-visual /pa/ pairing they would also report /pa/ with auditory presentation only. Nonetheless, the unimodal auditory experiment showed that the POA for unvoiced stimuli could not be dissociated, neither could it for /pa/. Thus, the use of auditory information alone should not result in a higher proportion of /pa/ responses. For the incongruent pairs, identification with the most positive SOA seems similar to unimodal unvoiced auditory perception, hence participants did not seem to use visual information, indicating that for this SOA integration did not take place. Similar results have been found by Vatakis and Spence (2007), who showed that judging simultaneity is more difficult when the gender of the speaker is congruent with the speech sound. Although there are also conflicting results, for speech the unity assumption seems plausible (Vroomen and Keetels, 2010).

One difference between simultaneity judgments and stimulus identification across SOAs seems to be that the point of maximal integration is more biased toward visual leading when explicitly asking about identity (Zampini et al., 2003; van Wassenhove et al., 2007). Therefore, varying SOAs and measuring unimodal perception might provide a different approach to measure whether integration occurs over a broader range of SOAs. This approach does not investigate whether two stimuli are perceived as simultaneously, but serves the goal to investigate the temporal patterns in which a unimodal stimulus influences the perception of another unimodal stimulus, for example the content of a stimulus. This judgment might be more natural, since in daily life, identifying stimuli is a more common act than explicitly judging their coincidence.

POSSIBLE NEURONAL MECHANISMS

Based on previous literature, the brain area most consistently involved in audio-visual integration is the posterior superior temporal sulcus (Calvert and Lewis, 2004). It has been found active during visual and audio-visual speech perception (Calvert et al., 1997; Callan et al., 2004), seems to be sensitive for congruent vs. incongruent speech signals (Calvert et al., 2000; van Atteveldt et al., 2004, 2010), and responds to audio-visual onset differences (van Atteveldt et al., 2007; Chandrasekaran and Ghazanfar, 2009). In the temporal domain it seems that different temporal features (co-variations between mouth velocity and speech envelope and visual-auditory speech onset differences) have to be combined to shape our percept. Chandrasekaran and Ghazanfar (2009) showed that different frequency bands are differently sensitive for faces and voices in superior temporal cortex. Although theta oscillations have been shown to be influenced by input from other senses (Lakatos et al., 2007; Kayser et al., 2008), they have not been shown to have specific effects dependent on the voice-face onset differences and might therefore mostly be used to parse the auditory signals, enhance auditory processing, and might even relate to the audio-visual TWI (Poeppel, 2003; Schroeder et al., 2008). However, higher frequency oscillations have been shown to vary dependent on voice-face onset differences, and might be involved in encoding the identity of a syllable, thus explaining the current results. This is consistent with the notion that the auditory speech system depends on theta as well as gamma frequencies (Poeppel, 2003), and this latter time-scale might also be important in coding differences in natural audio-visual onset differences, and its influence on perception. These temporal constraints however would have to be investigated, for example by using combined behavioral and electrophysiological measures, or using transcranial magnetic stimulation at varying time points.

CONCLUSION

Our findings show that within the integration window, visual information biases the auditory percept, specifically regarding the features in which the auditory signal is ambiguous (i.e., POA). Additionally, these findings indicate that natural temporal onset differences between auditory and visual input have a noteworthy influence on auditory perception. Although visual input has an influence over a wide temporal window during our experiment, we show that this initial binding of information does not

conclusively determine our percept. Instead, it serves as a prerequisite for other interaction processes to occur that eventually form our perceptual decision. The final percept is determined by the interplay between unimodal auditory and visual cues, along with natural audio-visual onset differences across syllables. These results shed light on the compositional nature of audio-visual speech, in which visual, auditory, and temporal onset cues are used to create a percept. This interplay of cues needs to be studied further to unravel the building blocks and neuronal basis of audio-visual speech perception.

ACKNOWLEDGMENTS

This study was supported by a grant from the Dutch Organization for Scientific Research (NWO; grant number 406-11-068).

REFERENCES

Andersen, T. S., Tiippana, K., and Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Brain Res. Cogn. Brain Res.* 21, 301–308. doi:10.1016/j.cogbrainres.2004.06.004

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

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. 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. doi:10.1044/1092-4388(2007/080)

Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41, 809–823. doi:10.1016/S0896-6273(04)00070-4

Bernstein, L. E., Auer, E. T., 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

Bien, N., ten Oever, S., Goebel, R., and Sack, A. T. (2012). The sound of size crossmodal binding in pitch-size synesthesia: a combined TMS, EEG and psychophysics study. *Neuroimage* 59, 663–672. doi:10.1016/j.neuroimage.2011.06.095

Blau, V., van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2008). Task-irrelevant visual letters interact with the processing of speech sounds in heteromodal and unimodal cortex. *Eur. J. Neurosci.* 28, 500–509. doi:10.1111/j.1460-9568.2008.06350.x

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., and Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *J. Cogn. Neurosci.* 16, 805–816. doi:10.1162/089892904970771

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 593–596. doi:10.1126/science.276.5312.593

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 Lewis, J. W. (2004). "Hemodynamic studies of audiovisual interactions," in *The Handbook of Multisensory Processes,* eds G. A. Calvert, C. Spence, and B. E. Stein (Cambridge: MIT Press), 483–502.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1001–1010. doi:10.1098/rstb.2007.2155

Chandrasekaran, C., and Ghazanfar, A. A. (2009). Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus. *J. Neurophysiol.* 101, 773–788. doi:10.1152/jn.90843.2008

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

Dixon, N. F., and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception* 9, 719–721. doi:10.1068/p090719

Doehrmann, O., and Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Res.* 1242, 136–150. doi:10.1016/j.brainres.2008.03.071

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

Ernst, M. O., and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn. Sci. (Regul. Ed.)* 8, 162–169. doi:10.1016/j.tics.2004.02.002

Gentilucci, M., and Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Exp. Brain Res.* 167, 66–75. doi:10.1007/s00221-005-0008-z

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., and Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* 25, 5004–5012. doi:10.1523/JNEUROSCI.0799-05.2005

Ghazanfar, A. A., Morrill, R. J., and Kayser, C. (2013). Monkeys are perceptually tuned to facial expressions that exhibit a theta-like speech rhythm. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1959–1963. doi: 10.1073/pnas.1214956110

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Front. Psychol.* 3:238. doi:10.3389/fpsyg.2012.00238

Grant, K. W., and Greenberg, S. (2001). "Speech intelligibility derived from asynchronous processing of auditory-visual information," in *Paper presented at the AVSP 2001-International Conference on Auditory-Visual Speech Processing*, (Washington, DC).

Grant, K. W., 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

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cereb. Cortex* 18, 1560–1574. doi:10.1093/cercor/bhm187

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

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* 49, 467. doi:10.1121/1.1912375

Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biol.* 8:e1000445. doi:10.1371/journal.pbio.1000445

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

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

Massaro, D. W. (1997). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am.* 100, 1777. doi:10.1121/1.417342

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

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.

Miller, J. L. (1977). Properties of feature detectors for VOT: the voiceless channel of analysis. *J. Acoust. Soc. Am.* 62, 641. doi:10.1121/ 1.381577

Müller, E., and MacLeod, G. (1982). Perioral biomechanics and its relation to labial motor control. *J. Acoust. Soc. Am.* 71, S33. doi:10.1121/1.2019340

Munhall, K., and Vatikiotis-Bateson, E. (1998). "The moving face during speech communication," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech,* eds R. Campbell, B. Dodd, and D. Burnham (Sussex: Taylor and Francis), 123–139.

Munhall, K., and Vatikiotis-Bateson, E. (2004). "Spatial and temporal constraints on audiovisual speech perception," in *The Handbook of Multisensory Processing,* eds G. A. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: The MIT Press), 177–188.

Pandey, P. C., Kunov, H., and Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *J. Aud. Res.* 26, 27–41.

Parise, C. V., and Spence, C. (2009). When birds of a feather flock together": synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE* 4:e5664. doi:10.1371/journal.pone.0005664

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

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci. (Regul. Ed.)* 12, 106–113. doi:10.1016/j.tics.2008.01.002

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

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

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., and Wallace, M. T. (2012). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. *Exp. Brain Res.* 219, 121–137. doi:10.1007/s00221-012-3072-1

Stevenson, R. A., and James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage* 44, 1210–1223. doi: 10.1016/j.neuroimage.2008.09. 034

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi:10.1121/1.1907384

Summerfield, A. (1987). "Some preliminaries to a theory of audiovisual speech processing,"in *Hearing by Eye II: The Psychology of Speechreading and Auditory-Visual Speech,* eds R. Campbell, B. Dodd, and D. Burnham (Hove: Erlbaum Associates), 58–82.

van Atteveldt, N., 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 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., Blau, V. C., Blomert, L., and Goebel, R. (2010). fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neurosci.* 11:11. doi:10.1186/1471-2202-11-11

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. doi:10.1073/pnas.0408949102

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., Maragos, P., Rodomagoulakis, I., and Spence, C. (2012). Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *Front. Integr. Neurosci.* 6:71. doi:10.3389/fnint.2012.00071

Vatakis, A., and Spence, C. (2006). 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. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Atten. Percept. Psychophys.* 69, 744–756. doi:10.3758/BF03193776

Vitkovitch, M., and Barber, P. (1994). Effect of video frame rate on subjects' ability to shadow one of two competing verbal passages. *J. Speech Lang. Hear. Res.* 37, 1204.

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

Wallace, M., Wilkinson, L., and Stein, B. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *J. Neurophysiol.* 76, 1246–1266.

Welch, R., and Warren, D. (1986). "Intersensory interactions," in *Handbook of Perception and Human Performance,* Vol. 1, eds K. Boff, L. Kaufmann, and J. Thomas (New York: Wiley), 25.21–25.36.

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* 88, 638. doi:10.1037/0033-2909.88.3.638

Wiener, F., and Miller, G. A. (1946). *Some Characteristics of Human Speech. Transmission and reception of sounds under combat conditions.* Summary Technical Report of Division 17 (Aalburg: National Defense Research Committee), 58–68.

Zampini, M., Shore, D. I., and Spence, C. (2003). Audiovisual temporal order judgments. *Exp. Brain Res.* 152, 198–210. doi:10.1007/s00221-003-1536-z

Zion Golumbic, E. M., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *J. Neurosci.* 33, 1417–1426. doi:10.1523/JNEUROSCI.3675-12.2013

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: 28 February 2013; paper pending published: 23 March 2013; accepted: 21 May 2013; published online: 26 June 2013.

*Citation: ten Oever S, Sack AT, Wheat KL, 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*

This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 16 July 2013
doi: 10.3389/fpsyg.2013.00432

Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life

Elena Kushnerenko[1]*, Przemyslaw Tomalski[1,2], Haiko Ballieux[1], Anita Potton[1], Deidre Birtles[1], Caroline Frostick[1] and Derek G. Moore[1]

[1] *Institute for Research in Child Development, School of Psychology, University of East London, London, UK*
[2] *Faculty of Psychology, University of Warsaw, Warsaw, Poland*

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
LouAnn Gerken, University of Arizona, USA
Marilyn Vihman, University of York, UK

****Correspondence:***
Elena Kushnerenko, Institute for Research in Child Development, School of Psychology, University of East London, Water Lane, London E15 4LZ, UK
e-mail: e.kushnerenko@gmail.com

The use of visual cues during the processing of audiovisual (AV) speech is known to be less efficient in children and adults with language difficulties and difficulties are known to be more prevalent in children from low-income populations. In the present study, we followed an economically diverse group of thirty-seven infants longitudinally from 6–9 months to 14–16 months of age. We used eye-tracking to examine whether individual differences in visual attention during AV processing of speech in 6–9 month old infants, particularly when processing congruent and incongruent auditory and visual speech cues, might be indicative of their later language development. Twenty-two of these 6–9 month old infants also participated in an event-related potential (ERP) AV task within the same experimental session. Language development was then followed-up at the age of 14–16 months, using two measures of language development, the Preschool Language Scale and the Oxford Communicative Development Inventory. The results show that those infants who were less efficient in auditory speech processing at the age of 6–9 months had lower receptive language scores at 14–16 months. A correlational analysis revealed that the pattern of face scanning and ERP responses to audiovisually incongruent stimuli at 6–9 months were both significantly associated with language development at 14–16 months. These findings add to the understanding of individual differences in neural signatures of AV processing and associated looking behavior in infants.

Keywords: audiovisual speech integration, infants' brain responses, ERPs, eye-tracking, language development, mismatch

INTRODUCTION

Visual speech cues are known to facilitate speech comprehension when auditory input is ambiguous, for example in a noisy environment, with the shape of the mouth partially indicating the sounds produced (Sumby and Pollack, 1954). Seeing someone speak may improve the comprehension of hard-to-understand passages even when hearing conditions are excellent (for a review see Campbell, 2008). A method for assessing capacities for audiovisual (AV) speech integration (AVSI) in adults and infants is to present simple video clips of people pronouncing syllables (/ba/ or /ga/) including clips where the visual and auditory speech components of the stimuli do not match (Kushnerenko et al., 2008). In these non-matching circumstances the fusion and the combination speech illusions may be perceived, a phenomenon known as the McGurk effect (McGurk and MacDonald, 1976). Of particular interest is what happens when a visual /ga/ and auditory /ba/ are presented together (VgaAba) as these are often fused by adults and perceived as the sound /da/ or /θa/. On the other hand a visual /ba/ dubbed onto auditory /ga/ (VbaAga) is often perceived as the combination /bga/.

Developmental studies of AVSI offer ambiguous results with respect to this phenomenon in infancy. Some behavioral studies indicate that infants as young as 4 months of age can perceive the McGurk "fusion" illusion (Rosenblum et al., 1997; Burnham and Dodd, 2004). Electrophysiological studies further indicate that 5 month-olds process the two kinds of audiovisually incongruent stimuli differently (Kushnerenko et al., 2008), suggesting that these lead to the same "combination" and "fusion" effects as are seen in adults. In this study, the AV mismatch response (AVMMR) was recorded in response to the VbaAga-combination condition but not to the VgaAba-fusion.

On the other hand, Desjardins and Werker (2004) demonstrated that AV integration is not an obligatory process in young infants and that it may require a degree of experience with language before emerging. Further, Massaro (1984) hypothesized that differences between adults and children in AVSI can be explained by different levels of attention to the visual component of the stimuli. For example, the use of visual cues during AV processing of speech is known to be less efficient in children and adults with language-learning disabilities (Norrix et al., 2006, 2007). Also, difficulties in integrating auditory and visual aspects of speech perception have been reported in children with specific language impairment (Pons et al., 2013) and in autism spectrum disorder (ASD; Guiraud et al., 2012; Megnin et al., 2012).

Attention to visual speech cues appears to undergo significant changes over the first year of life. Lewkowicz and Hansen-Tift (2012) demonstrated a developmental shift in visual attention to articulating faces within the first 12 months of life from an initial tendency to look at the eyes rather than the mouth, followed by a marked increase in looking at the mouth, returning to preference for the eyes at 12 months of age. This pattern in attentional shifts may correspond with transitional periods in speech acquisition in infancy. For example, recent studies have demonstrated that visual attention to the eye region at 6 months, but not at 9 and 12 months, is associated with better social and communicative outcomes at the age of 18 months (Schietecatte et al., 2012; Wagner et al., 2013).

Visual attention, specifically during AVSI, has recently been investigated in detail in 6- to 9-month-old infants using the paradigm developed by Kushnerenko, Tomalski, and colleagues (Tomalski et al., 2012; Kushnerenko et al., 2013). In this eye-tracking (ET) paradigm, faces articulating either /ba/ or /ga/ syllables were displayed along with the original auditory syllable (congruent VbaAba and VgaAga), or a mismatched one (incongruent VbaAga and VgaAba). By measuring the amount of looking to the eyes and mouth of articulating faces, it was found that younger infants (6–7 months) may not perceive mismatching auditory /ga/ and visual /ba/ (VbaAga) cues in the same way as adults, that is, as the combination /bga/ (McGurk and MacDonald, 1976) but process these stimuli as a mismatch between separate cues and "reject" them as a source of unreliable information, and therefore allocate less attention to them. Using the same stimuli, Kushnerenko et al. (2013) also found that the AVMMR brain response to these stimuli showed large individual differences between 6 and 9 months of age, and that these differences were strongly associated with differences in patterns of looking to the speaker's mouth. Interestingly, the amplitude of the AVMMR was inversely correlated with looking time to the mouth, which is consistent with the results found by Wagner et al. (2013). These results suggest that at this age sufficient looking toward the eyes may play a pivotal role for later communicative and language development. Given these results, and the fact that infants as young as 2–5 months of age are able to match auditory and visual speech cues (Kuhl and Meltzoff, 1982; Patterson and Werker, 2003; Kushnerenko et al., 2008; Bristow et al., 2009), we hypothesized that individual differences in visual attention and brain processing of AV speech sounds should predict language development at a later age.

In the current paper we report the results of a follow-up study with infants who at the age of 6- to 9-months completed an AVSI task with matching and mismatching speech cues. AVSI was assessed with both ET and event-related potential (ERP) measures in the same task, reported elsewhere (Tomalski et al., 2012; Kushnerenko et al., 2013). For the present follow-up, infants attended a session when they were 14- to 16-months-old, and their early language and communicative development was assessed using language assessment tests. The sample had been recruited from areas with a high multiple deprivation index with the purpose of recruiting a diverse sample in terms of family socio-economic status (SES) in order to capture a range of abilities. Several studies have indicated that children from low-SES areas have weaker language skills at preschool age (Raizada et al., 2008) and deficits in selective attention related to speech processing, including a reduced ability to filter out irrelevant auditory information (Stevens et al., 2009). We therefore expected a representative proportion of our sample of infants to be at risk of later language related difficulties.

There is now evidence for the existence of early individual differences in how young infants visually scan social stimuli (Kushnerenko et al., 2013). There is also evidence that these individual differences can be predictive of later language (e.g., Young et al., 2009) and communicative development (Wagner et al., 2013). Also, auditory-only speech sound discrimination in 6-month-olds predicts later vocabulary (e.g., Tsao et al., 2004). Given this evidence we have sought to establish whether individual differences in AV speech processing at 6- to 9-months of age predict language development at 14- to 16 months. In particular we measured the neural responses and the amount of time spent fixating the eyes and the mouth of articulating faces with mismatching AV speech cues. We hypothesized that the pattern of visual attention to incongruent AV speech cues in infancy and sensitivity to AV mismatch as reflected by brain responses might be a significant predictor of receptive and expressive language in toddlers.

MATERIALS AND METHODS

PARTICIPANTS

All 37 infants had previously participated in an ET AV task (Tomalski et al., 2012) when aged between 6 and 9 months (10 were boys; the mean age was 33.5 weeks, SD = 2.8 weeks). Twenty-two of these infants (6 boys, mean age 30.7 weeks, SD = 4.3 weeks) also participated in an ERP AV task (Kushnerenko et al., 2013). The birth weight of infants and gestational ages were in the normal range (mean weight 3377.6 g; mean gestational age 39.59 weeks). The average total income of the families was £52,401 and ranged from £4,800 to £192,000, which represents a large income range (see **Table 1**). The age range for this study was chosen because neural signatures of auditory processing demonstrate different rates of maturation during this age period, with some 6 month-olds showing a more mature ERP pattern and some 9 month-olds a less mature one (Kushnerenko et al., 2002). The study was approved by the local ethics committee and conformed to the Declaration of Helsinki. Prior to the study parents gave written informed consent for their child's participation.

LANGUAGE ASSESSMENT AT 14–16 MONTHS

Infants were assessed individually using the PreSchool Language Scale-4 (PLS-4; Zimmerman et al., 2002) between 14 and 16 months (mean = 14.7, SD = 0.7). The PLS-4 is a norm-referenced test of receptive and expressive language ability for ages from birth to 6 years and 11 months. The test consists of a picture book and manipulative toys designed to engage a child in order to elicit responses to test items. The test gives two standardized sub-scales, auditory comprehension (AC) and expressive communication (EC), and a total score. During the follow-up parents were also asked to complete the Oxford Communicative Development Inventory (OCDI, a UK adaptation of the MacArthur-Bates CDI). The OCDI is a tool for assessing the development of receptive and productive vocabulary through parental report and is typically

Table 1 | Demographic characteristics of the higher AC-PLS and lower AC-PLS groups of infants (standard deviation).

Measure		All infants ($n = 37$)	Lower AC-PLS ($n = 19$)	Higher AC-PLS ($n = 18$)
Age at second session (weeks)		65.49 (3.24)	65.05 (2.62)	65.65 (3.67)
Gender	Female Male	27 10	13 6	14 4
Gestational age (weeks)		39.59 (1.87)	40.00 (1.81)	39.24 (1.89)
Birth weight (grams)		3377.6 (413.9)	3400.17 (366.9)	3358.33 (458.47)
Average income (£)		52,401 (43,062)	43,002 (35,901)	60,518 (47,732)
Mother SOC	(1)	47.5%	50.0%	63.6%
	(2)	17.5%	22.2%	13.6%
	(3)	25.0%	27.7%	22.7%

used with children aged from about 11–26 months (Hamilton et al., 2000). Basic demographic information on family income, parental education and occupation was collected from the primary caregivers (see **Table 1**) via a study-designed questionnaire (Tomalski et al., 2013).

EYE-TRACKING TASK AT 6–9 MONTHS

Infants were seated on their caregiver's lap in a dimly lit room. They were seated approximately 60 cm in front of a Tobii T120 eye-tracker monitor (17″ diameter, screen refresh rate 60 Hz, ET sampling rate of 120 Hz, spatial accuracy 0.5°). Prior to the experiment each infant's eye movements were calibrated using a five-point routine in order to ensure positional validity of gaze measurements. At least 50% of samples were recorded from each infant during each trial. The parent's view of the stimulus monitor was obscured to prevent interference with the infant's looking behavior. Eye movements were monitored continuously during each recording. Every infant observed a total of ten trials. Before each trial, infants' attention was directed to the screen by colorful animations with sound, and these were terminated as soon as the infant fixated them. For more details on the ET task see Tomalski et al. (2012).

The stimuli were two video clips of female native English speakers articulating /ba/ and /ga/ syllables and two incongruent pairs which were created from the original AV stimuli by dubbing the auditory /ba/ onto a visual /ga/ (VgaAba) and vice versa (VbaAga). Sound onset in each clip was 360 ms from stimulus onset, and auditory syllable duration was 280–320 ms. The total duration of one AV stimulus was 760 ms. For more information on the stimuli see Kushnerenko et al. (2008). Each trial contained 10 repetitions of one type of stimulus and the trial duration was 7600 ms (760 ms × 10). The entire sequence lasted approximately 2 min.

EVENT-RELATED POTENTIAL STUDY AT 6–9 MONTHS

The paradigm and stimuli for this task were the same as in Kushnerenko et al. (2008).

The same AV stimuli as in the ET study were presented in a pseudorandom order. Videos were displayed on a CRT monitor (30 cm diameter, 60 Hz refresh rate) with a black background. The infants were seated on the caregiver's lap in an acoustically and electrically shielded booth. They were seated at a distance of 80 cm from the monitor. At that distance the faces on the monitor were approximately life size. Sounds were presented at about a 65 dB level via two loudspeakers behind the screen. The recording time varied from 4 to 6 min, depending on each infant's attention to the stimuli. The behavior of the infants was videotaped and coded off-line for electroencephalography (EEG) artifact rejection.

High-density EEG was recorded with a 128-channel Hydrocel Sensor Net (EGI Inc.) referenced to the vertex (Tucker, 1993). The EEG signal was amplified, digitized at 500 Hz, and band-pass filtered from 0.1 to 200 Hz. The signal was off-line low-pass filtered at 30 Hz and segmented into epochs starting 100 ms before and ending 1,000 ms after the AV stimulus onset. Channels contaminated by eye or motion artifacts were rejected manually, and trials with more than 20 bad channels were excluded. In addition, video recordings of the infants' behavior were coded frame-by-frame, and trials during which the infant did not attend to the face were excluded from further analysis. Following artifact rejection, the average number of trials for an individual infant accepted for further analysis was 37.4 for /ba/, 36.7 for /ga/, 37.6 for VgaAba, and 37.8 for VbaAga. Although uncommon for adult ERP studies, this number of accepted trials has proven to be sufficient in infant studies (Dehaene-Lambertz and Dehaene, 1994; Friederici et al., 2007; Kushnerenko et al., 2008; Bristow et al., 2009; Guiraud et al., 2011).

Artifact-free segments were re-referenced to the average reference and then averaged for each infant within each condition. A baseline correction was performed by subtracting mean amplitudes in the 260–360 ms window from the video onset (i.e., immediately before the sound onset) to minimize the effects of any ongoing processing from the preceding stimulus. For the statistical analyses we bilaterally defined channel groups: frontal (area between Fp1, F3, and Fz on the left and symmetrical on the right), central (area between F3, C3, and Cz on the left and symmetrical on the right), occipital (area between O1, P3, and Pz on the left and symmetrical on the right) and temporo-parietal (covering area between P3 and left mastoid and P4 and the right mastoid). The analyses were conducted on mean amplitudes within the time window between 290 and 390 ms from the sound onset for AVMMR (Kushnerenko et al., 2008) and between 140 to 240 ms from the sound onset for infantile P2 (Kushnerenko et al., 2007). The correlation analysis was performed for the frontal and central ERP mean amplitudes and looking time to the eyes and mouth as a percentage of total looking time to the face in both audiovisually mismatching conditions VbaAga and VgaAba. Partial correlations controlled for the age at the first session, total family income, and maternal occupation. The last two variables were taken as indicators of SES of the family, and have been previously found to be associated with the power of frontal gamma oscillations (Tomalski et al., 2013).

RESULTS

Pearson correlations were computed in order to determine whether neural or behavioral signatures of AV processing at 6–9 months, specifically the processing of a mismatch between auditory and visual speech cues, is associated with language outcome at 14–16 months of age. In this analysis we partialled out age at first assessment, total family income, and maternal occupation. These factors are known to contribute to individual differences in language outcomes, and we wanted to examine how well early AVSI responses can predict language outcomes, having controlled for these potential mediating variables.

ASSOCIATIONS BETWEEN ATTENTION TO AUDIOVISUAL SPEECH AT 6–9 MONTHS AND LANGUAGE DEVELOPMENT AT 14–16 MONTHS

Partial Pearson correlations confirmed that PLS-4 AC scores were significantly negatively correlated with looking time to the mouth in the VbaAga condition (**Table 2**), and positively correlated with looking time to the eyes in the VgaAba condition (see also **Figure 1**). These results indicate a similar tendency for both incongruent AV conditions: infants who received higher scores for their language development had shorter looking times to the mouth area and/or longer looking times to the eyes when they encountered AV mismatch.

As **Figure 1** demonstrates, the longer infants looked at the eyes in the VgaAba condition, the better their AC was 1 year later. In addition, there were significant correlations between looking time to the eyes in this condition and the Oxford CDI productive vocabulary (OCDI) score (partial-$r = 0.42$, $p = 0.01$), as well as a marginally significant association with OCDI comprehension score (partial-$r = 0.32$, $p = 0.06$).

ASSOCIATIONS BETWEEN ERP MEASURES OF AV PROCESSING AT 6–9 MONTHS AND LANGUAGE DEVELOPMENT AT 14–16 MONTHS

Given the result that lower language scores at 14–16 months were associated with longer looking time to the mouth area, we also expected an association with larger frontal P2 amplitudes in response to the VgaAba-fusion condition. In a previous study we have found an association between looking time to the mouth and frontal P2 amplitude (Kushnerenko et al., 2013). Indeed, partial correlation coefficients were significant (partial-$r = -0.68$, $p = 0.001$, partial-$r = 0.48$, $p = 0.04$) for PLS-4 AC scores and the amplitude of the infantile P2 over frontal areas in response to the same stimulus (VgaAba; see **Figure 2**). It should be noted that correlations for the mean voltage over the frontal area were negative, which indicates that larger P2 amplitudes are associated with poorer language comprehension.

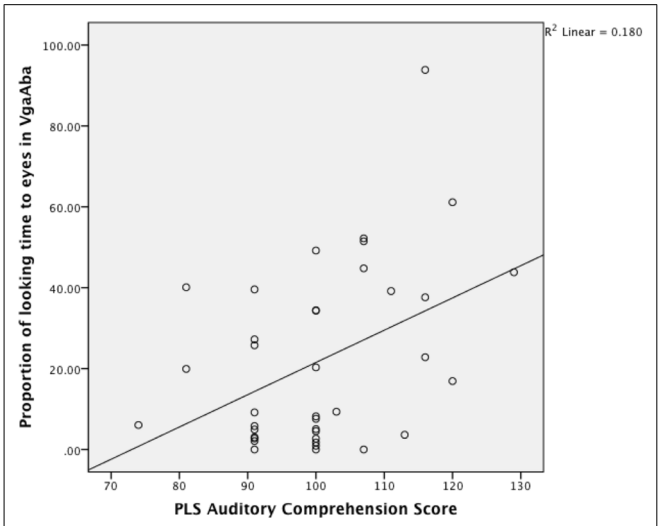


FIGURE 1 | A scatterplot showing the relationship between looking time to eyes in VgaAba condition at 6–9 months of age and PLS auditory comprehension scale at 14–16 months.

For illustration purposes, the participants were median split into low and high groups on the basis of AC (see **Table 1** for the demographics profile of these two groups). Note that although they appear to differ in income, there were no significant differences between these groups on demographic measures.

Figure 3 demonstrates that while the P2 amplitude in response to congruent AV /ba/ and /ga/ stimuli is of about the same

Table 2 | Partial correlations for PLS-4 and Oxford CDI scores at 14–16 months and eye-tracking and ERP measurements at 6–9 months of age (partial-*r* and p).

	Looking time to eyes in VbaAga	Looking time to mouth in VbaAga	Looking time to eyes in VgaAba	Looking time to mouth in VgaAba	Frontal left P2 amplitude	Frontal right P2 amplitude
Oxford CDI comprehension	0.09	0.01	0.32	−0.27	−0.06	−0.10
	0.59	0.94	0.05	0.12	0.81	0.66
Oxford CDI production	0.01	−0.19	0.41	−0.29	−0.18	−0.04
	0.96	0.26	0.01 *	0.09	0.46	0.88
PLS auditory comprehension	0.31	−0.34	0.35	−0.16	−0.68	−0.48
	0.07	0.04*	0.03*	0.35	0.001*	0.04*
PLS expressive communication	−0.20	0.05	−0.15	0.03	−0.09	−0.16
	0.25	0.74	0.37	0.87	0.74	0.55

*$p < 0.05$.

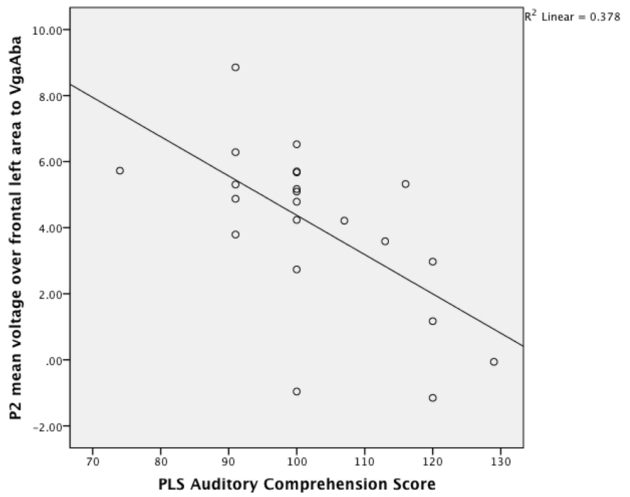


FIGURE 2 | A scatterplot showing the relationship between the mean P2 voltage over the left frontal area (140–240 ms) in response to VgaAba at 6–9 months of age and PLS auditory comprehension scale at 14–16 months.

amplitude in both groups, in response to incongruent VgaAba stimuli it appears to be larger over the frontal area in infants with lower AC-PLS scores (F3 channel). In addition, although no significant associations were found between language outcome and the amplitude of the AVMMR, this brain response to incongruent VbaAga was only observed in the higher AC-PLS group of infants.

Figure 4 shows the percentage of looking time to the eyes and mouth in both groups of infants. The subgroup of infants with higher AC-PLS scores showed generally longer looking times to the eyes and shorter looking times to the mouth. However, the difference between groups was significant only for the incongruent conditions (two-sample t test, $p < 0.03$ for VbaAga-Mouth, and $p < 0.05$ for VgaAba-Eyes).

DISCUSSION

The aim of the present study was to investigate whether individual differences in neural and behavioral markers of AV processing in infancy might be indicative of later language development. The follow-up assessment revealed significant associations between ET and EEG measures at 6–9 months and language development measures at 14–16 months of age. Specifically, infants who spent a longer time watching the eyes of a female face when auditory and visual speech cues did not match performed better on the AC scale of the Preschool Language Scale (PLS-4). The level of functioning on the vocabulary scale of the Oxford CDI was also consistent with this result and showed a significant association with looking time to the eyes in the mismatched condition. In addition, infants who had higher AC-PLS scores appeared to look less at the

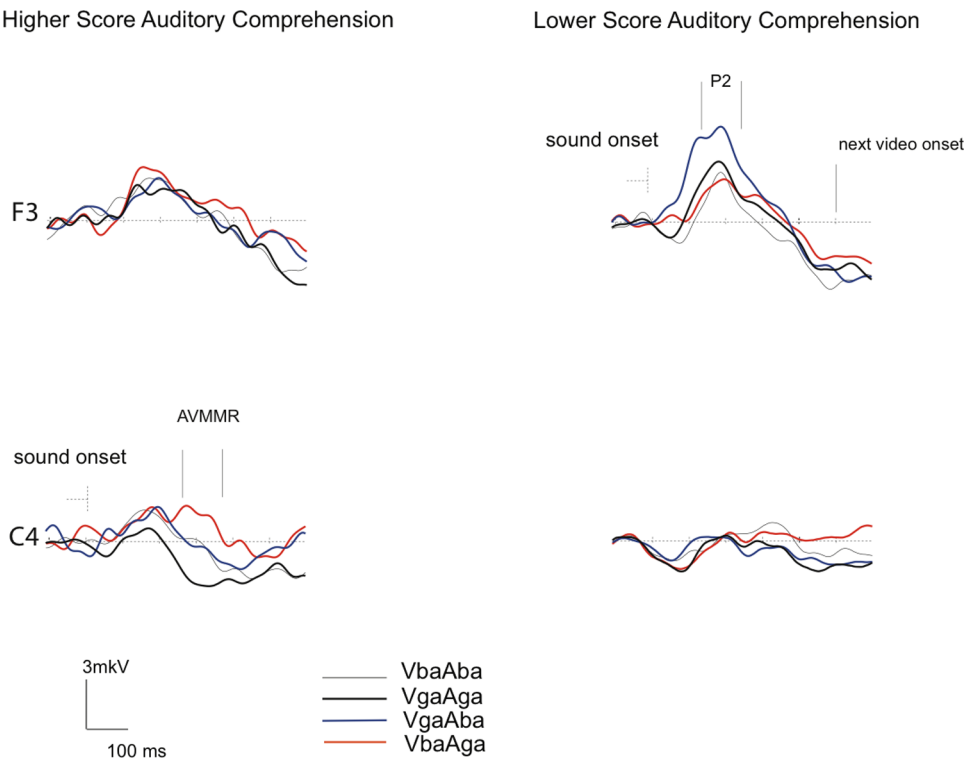


FIGURE 3 | Grand-averaged ERP responses to congruent and incongruent AV stimuli in 6–9 month-old infants with higher and lower AC-PLS scores at the follow-up age. ERP responses plotted for VbaAba (thin grey), VbaAga (orange), VgaAba (blue), and VgaAga (black) time-locked to the sound onset. Selected channels are shown according to 10–20 system.

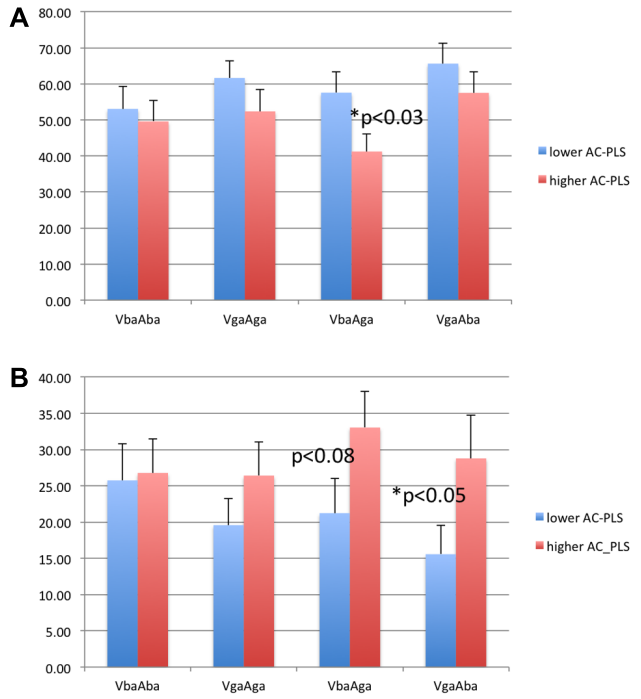


FIGURE 4 | Looking time to the mouth (A) and eyes (B) as percentage of looking time to the face in 6–9 month-old infants with higher and lower AC-PLS scores at the follow-up age.

mouth during presentation of the saliently mismatched condition (VbaAga).

As shown previously, this pattern of responses where infants show shorter looking times to the mouth when auditory and visual stimuli are mismatched is a transitional phase in development between the ages of 6 and 9 months. In an earlier study Tomalski et al. (2012) found a positive association between age and the amount of looking at the mouth for mismatched AV speech cues.

By the time infants reach 9 months of age, they clearly show longer looking to the mouth while watching incongruent AV speech cues compared with the congruent syllables (Tomalski et al., 2012). At the same time the neural AVMMR in response to AV mismatch significantly decreases in amplitude, indicating that this signature of AV mismatch processing in infancy is transitory (Kushnerenko et al., 2013).

Nardini et al. (2010) proposed that not integrating sensory cues might be adaptive for younger children because they must learn not only to combine cues but also to establish whether these cues are reliable, and whether some cues should be ignored. The developmental pattern observed between 6 and 9 months of age in ET (Tomalski et al., 2012) is consistent with this idea: shorter looking times to the mouth in the mismatched condition indicate that younger infants ignored unreliable and confusing visual cues.

The results of the present study indicate that the ability to detect a mismatch between visual and auditory cues during this transitional phase might be indicative of AC in the second year of life. This may imply that infants who spent longer time watching mouth articulation during mismatched AV trials might have difficulties with recognizing the auditory component and therefore might seek more information from lip movements. By contrast, toddlers with higher AC scores may have correctly recognized the speech sound during the ET and ERP tasks at 6–9 months, but did not find it helpful to attend to the distracting and unreliable lip movements. Thus, longer looking times to the eyes during mismatched AV trials in these infants may indicate that they were searching for additional social cues to resolve the ambiguity of these stimuli.

This assumption is consistent with a recent study that shows visual attention to the eye region (measured using ET at 6 months) to be associated with better social outcome at the age of 18 months, as measured by the Communication and Symbolic Behavior Scales (Wagner et al., 2013; see also Schietecatte et al., 2012). Interestingly, the association was significant for younger infants (6-month-olds) but not for 9- and 12-month-olds. This finding illustrates once again that the pattern of visual attention in infants largely depends on their maturational level (Lewkowicz and Hansen-Tift, 2012; Kushnerenko et al., 2013).

On the other hand, another longitudinal study yielded the opposite pattern of results: infants with longer fixation on the mouth demonstrated better expressive language skills later on (Young et al., 2009). However, the group of infants tested in the study by Young and colleagues had a higher familial risk of ASD and the design of the study was different, with infants only seeing congruent live mother–infant interaction and no confusing AV information. In the present study, differences in looking behavior between infants with higher and lower AC scores were only significant for incongruent AV conditions (VbaAga and VbaAga). Therefore, the results of Young and colleagues (2009) seem to demonstrate a different phenomenon and are not comparable with those of the current study. In addition, in the study of Young and colleagues (2009) the correlations were found for the expressive language score and not for AC. We propose therefore that attention to the mouth is more important for the development of expressive language because it facilitates imitation and is useful for learning how to articulate particular speech sounds (Howard and Messum, 2011). On the other hand, AC abilities are likely to be more related to the accuracy of auditory processing in young infants. Attention to the eyes then may assist in learning new object labels. Infants increasingly use referential gaze as a cue to direct their looking toward an object that is being named (e.g., Gliga and Csibra, 2009) and benefit from referential gaze in their language learning (Houston-Price et al., 2006).

In the present study, significant associations were also found between receptive language score at 14–16 months and ERP measures of AV processing at 6–9 months of age. A larger amplitude of the frontal P2 was found in response to the incongruent VgaAba stimulus in a subgroup of infants with lower AC score at the follow-up age. Larger P2 amplitudes (positive over frontal and negative over occipital areas) to incongruent AV stimuli have previously been observed in infants who spent longer time attending to lip articulations than to eyes (Kushnerenko et al., 2010, 2013).The increased P2 may have contributions from the activity of visual

areas, therefore demonstrating that infants who look longer at the mouth might be processing visual cues more intensively than auditory ones. In the present follow-up study, both the increased frontal P2 amplitude and longer looking time to the mouth during the mismatch VbaAga condition in infancy were associated with less advanced AC later in development. One possible explanation for this could be that infants who have less accurate or less mature auditory speech processing at the age of 6–9 months rely more on using visual cues when ambiguous speech stimuli are presented. This pattern of results may indicate that a visual-over-auditory bias in sensory processing of speech cues at 6 months of age can be predictive of less advanced auditory speech comprehension at the age of 14–16 months.

To summarize, in the present study the larger frontal P2 amplitudes to the ambiguous AV stimuli were associated with lower AC scores on language scales in 14–16 month-old toddlers. In addition, there was a significant association between longer looking times to the eyes than to the mouth in the incongruent conditions and the higher AC score (and the opposite tendency for longer looking times to the mouth). These findings provide important evidence that early markers of infants' visual attention relate not only to their social development (Schietecatte et al., 2012; Wagner et al., 2013) but also to their later language development. The current results also demonstrate that early electrophysiological indices of AV speech processing are indicative of language comprehension in the second year of life.

ACKNOWLEDGMENTS

We acknowledge the financial support of Eranda Foundation, and the University of East London (Promising Researcher Grant to Elena Kushnerenko and School of Psychology funding for Przemyslaw Tomalski and Derek G. Moore). We thank all families for their participation in the study.

REFERENCES

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., et al. (2009). Hearing faces: how the infant brain matches the face it sees with the speech it hears. *J. Cogn. Neurosci.* 21, 905–921. doi: 10.1162/jocn.2009.21076

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

Campbell, R. C.-P. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1001–1010. doi: 10.1098/rstb.2007.2155

Dehaene-Lambertz, G., and Dehaene, S. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature* 28, 293–294. doi: 10.1038/370292a0

Desjardins, R. N., and Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Dev. Psychobiol.* 45, 187–203. doi: 10.1002/dev.20033

Friederici, A. D., Friedrich, M., and Christophe, A. (2007). Brain responses in 4-month-old infants are already language specific. *Curr. Biol.* 17, 1208–1211. doi: 10.1016/j.cub.2007.06.011

Gliga, T., and Csibra, G. (2009). One-year-old infants appreciate the referential nature of deictic gestures and words. *Psychol. Sci.* 20, 347–353. doi: 10.1111/j.1467-9280.2009.02295.x

Guiraud, J. A., Kushnerenko, E., Tomalski, P., Davies, K., Ribeiro, H., Johnson, M. H., et al. (2011). Differential habituation to repeated sounds in infants at high risk for autism. *Neuroreport* 22, 845–849. doi: 10.1097/WNR.0b013e32834c0bec

Guiraud, J. A., Tomalski, P., Kushnerenko, E., Ribeiro, H., Davies, K., Charman, T., et al. (2012). Atypical audiovisual speech integration in infants at risk for autism. *PLoS ONE* 7:e36428.doi: 10.1371/journal.pone.0036428

Hamilton, A., Plunkett, K., and Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory: lower scores in the UK than the USA. *J. Child Lang.* 27, 689–705.

Houston-Price, C., Plunkett, K., and Duffy, H. (2006). The use of social and salience cues in early word learning. *J. Exp. Child Psychol.* 95, 27–55. doi: 10.1016/j.jecp.2006.03.006

Howard, I. S., and Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition learning to pronounce. *Motor Control* 1, 85–117.

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

Kushnerenko, E., Čeponiené, R., Balan, P., Fellman, V., Näätänen, R., and Huotilainen, M. (2002). Maturation of the auditory change-detection response in infants: a longitudinal ERP study. *Neuroreport* 13, 1843–1848.

Kushnerenko, E., Teinonen, T., Volein, A., and Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11442–11445. doi: 10.1073/pnas.0804275105

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., et al. (2010). Audiovisual speech integration: visual attention to articulation affects brain responses in 6-9 month old infants. *Paper presented at EPS/SEPEX,* 15–17 April 2010, Granada, Spain. doi: 10.1016/j.wocn.2009.04.002

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., et al. (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *Eur. J. Neurosci.* doi: 10.1111/ejn.12317

Kushnerenko, E., Winkler, I., Horváth, J., Näätänen, R., Pavlov, I., Fellman, V., et al. (2007). Processing acoustic change and novelty in newborn infants. *Eur. J. Neurosci.* 26, 265–274. doi: 10.1111/j.1460-9568.2007.05628.x

Lewkowicz, D. J., and Hansen-Tift, A. M. C.-P. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1431–1436. doi: 10.1073/pnas.1114783109

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Dev.* 55, 1777–1788. doi: 10.2307/1129925

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

Megnin, O., Flitton, A., Jones, C. R. G., De Haan, M., Baldeweg, T., and Charman, T. (2012). Audiovisual speech integration in autism spectrum disorders: ERP evidence for atypicalities in lexical-semantic processing. *Autism Res.* 5, 39–48. doi: 10.1002/aur.231

Nardini, M., Bedford, R., and Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17041–17046. doi: 10.1073/pnas.1001699107

Norrix, L. W., Plante, E., and Vance, R. (2006). Auditory-visual speech integration by adults with and without language-learning disabilities. *J. Commun. Disord.* 39, 22–36. doi: 10.1016/j.jcomdis.2005.05.003

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)

Patterson, M. L., and Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Dev. Sci.* 6, 191–196. doi: 10.1111/1467-7687.00271

Pons, F., Andreu, L., Sanz-Torrent, M., Buil-Legaz, L., and Lewkowicz, D. J. (2013). Perception of audiovisual speech synchrony in Spanish-speaking children with and without specific language impairment. *J. Child Lang.* 40, 687–700. doi: 10.1017/S0305000912000189

Raizada, R. D. S., Richards, T. L., Meltzoff, A., and Kuhl, P. K. (2008). Socioeconomic status predicts hemispheric specialisation of the left inferior frontal gyrus in young children. *Neuroimage* 40, 1392–1401. doi: 10.1016/j.neuroimage.2008.01.021

Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (1997). The McGurk effect in infants. *Percept. Psychophys.* 59, 347–357. doi: 10.3758/BF03211902

Schietecatte, I., Roeyers, H., and Warreyn, P. (2012). Can infants' orientation to social stimuli predict

later joint attention skills? Br. *J. Dev. Psychol.* 30, 267–282. doi: 10.1111/j.2044-835X.2011.02039.x

Stevens, C., Lauinger, B., and Neville, H. (2009). Differences in the neural mechanisms of selective attention in children from different socioeconomic backgrounds: an event-related brain potential study. *Dev. Sci.* 12, 634–646. doi: 10.1111/j.1467-7687.2009.00807.x

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

Tomalski, P., Moore, D. G., Ribeiro, H., Axelsson, E. L., Murphy, E. l., Karmiloff-Smith, A., et al. (2013). Socio-economic status and functional brain development – associations in early infancy. *Dev. Sci.* doi: 10.1111/desc.12079

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., et al. (2012). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *Eur. J. Dev. Psychol.* 1–14. doi: 10.1080/17405629.2012.728076

Tsao, F.-M., Liu, H.-M., and Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Dev.* 75, 1067–1084. doi: 10.1111/j.1467-8624.2004.00726.x

Tucker, D. M. (1993). Spatial sampling of head electrical fields: the geodesic sensor net. *Electroencephalogr. Clin. Neurophysiol.* 87, 154–163.

Wagner, J. B., Luyster, R. J., Yim, J. Y., Tager-Flusberg, H., and Nelson, C. A. (2013). The role of early visual attention in social development. *Int. J. Behav. Dev.* 37, 118–124. doi: 10.1177/0165025412486064

Young, G. S., Merin, N., Rogers, S. J., and Ozonoff, S. (2009). Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Dev. Sci.* 12, 798–814. doi: 10.1111/j.1467-7687.2009.00833.x

Zimmerman, I., Steiner, V., and Pond, R. (2002). *Preschool Language Scale*, 4th Edn. San Antonio: The Psychological Corporation.

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 April 2013; paper pending published: 03 May 2013; accepted: 23 June 2013; published online: 16 July 2013.

Citation: Kushnerenko E, Tomalski P, Ballieux H, Potton A, Birtles D, Frostick C and Moore DG (2013) Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life. Front. Psychol. ***4****:432. doi: 10.3389/fpsyg.2013.00432*

This article was submitted to Frontiers in Language Sciences, a specialty of Frontiers in Psychology.

frontiers in
PSYCHOLOGY

GENERAL COMMENTARY
published: 24 March 2014
doi: 10.3389/fpsyg.2014.00257

Multisensory integration, learning, and the predictive coding hypothesis

Nicholas Altieri*

ISU Multimodal Language Processing Lab, Department of Communication Sciences and Disorders, Idaho State University, Pocatello, Idaho, USA
**Correspondence: altinich@isu.edu*

Edited by:
Albert Costa, University Pompeu Fabra, Spain

Reviewed by:
Ryan A. Stevenson, University of Toronto, Canada
Jordi Navarra, Fundació Sant Joan de Déu - Parc Sanitari Sant Joan de Déu - Hospital Sant Joan de Déu, Spain

Keywords: predictive coding, Bayesian inference, audiovisual speech integration, EEG, parallel models

A commentary on

Speech through ears and eyes: interfacing the senses with the supramodal brain
by van Wassenhove, V. (2013). Front. Psychol. 4:388. doi: 10.3389/fpsyg.2013.00388

The multimodal nature of perception has generated several questions of importance pertaining to the encoding, learning, and retrieval of linguistic representations (e.g., Summerfield, 1987; Altieri et al., 2011; van Wassenhove, 2013). Historically, many theoretical accounts of speech perception have been driven by descriptions of auditory encoding; this makes sense because normal-hearing listeners rely predominantly on the auditory signal. However, from both evolutionary and empirical standpoints, comprehensive neurobiological accounts of speech perception must account for interactions across sensory modalities and the interplay of cross-modal and articulatory representations. These include auditory, visual, and somatosensory modalities.

In a recent review, van Wassenhove (2013) discussed key frameworks describing how visual cues interface with the auditory modality to improve auditory recognition (Sumby and Pollack, 1954), or otherwise contribute to an illusory percept for mismatched auditory-visual syllables (McGurk and MacDonald, 1976). These frameworks encompass multiple levels of analysis. Some of these higher cognitive processing models that discuss parallel processing (Altieri and Townsend, 2011) or the independent extraction of features from the auditory and visual modalities (Massaro, 1987, Fuzzy Logical Model of Perception), early feature encoding (van Wassenhove et al., 2005), and encoding/timing at the neural level (Poeppel et al., 2008; Schroeder et al., 2008).

This commentary on van Wassenhove (2013) will examine predictive coding hypotheses as one theory for how visemes are matched with auditory cues. Crucially, a hypothesized role shall be emphasized for cross-modal neural plasticity and multisensory learning in reinforcing the sharing of cues across modalities into adulthood.

PREDICTIVE ENCODING AND FIXED PRIORS

A critical question in speech research concerns how time-variable signals interface with internal representations to yield a stable percept. Although speech signals are highly variable (multiple talkers, dialects, etc.), our percepts appear stable due to dimensionality reduction. These questions become even more complex in multisensory speech perception since we are now dealing with the issue of how visual speech gestures coalesce with the auditory signal as the respective signals unfold at different rates and reach cortical areas at different times. In fact, these signals must co-occur within an optimal spatio-temporal window to have a significant probability of undergoing integration (Conrey and Pisoni, 2006; Stevenson et al., 2012).

The *predictive coding hypothesis* incorporates these aforementioned observations to describe integration in the following ways: (1) Temporally congruent auditory and visual inputs will be processed by cortical integration circuitry, (2), internal representations ("fixed Bayesian priors") are compared and matched against the inputs, and (3) hypotheses about the intended utterance are actively generated. van Wassenhove et al.'s (2005) EEG study exemplified key components of the visual predicative coding hypothesis. When presented with auditory and visual syllables in normal conversational settings, the visual signal leads the auditory by tens or even hundreds of milliseconds. Thus, featural information in the visual signal constrains predictions about the content of the auditory signal. The authors showed that early visual speech information speeds-up auditory processing, as evidenced by temporal facilitation in the early auditory ERPs. This finding was interpreted as a reduction in the residual error in the auditory signal by the visual signal. One promising hypothesis is that visual information interacts with the auditory cortex in such a way that it modulates excitability in auditory regions via oscillatory phase resetting (Schroeder et al., 2008). Predictive coding hypotheses may also be extended to account for broad classes of stimuli including speech and non-speech, and matched and mismatched signals—all of which have been shown to evoke early ERPs associated with visual prediction (Stekelenburg and Vroomen, 2007).

FIXED PRIORS

Hypothetically, visual cues can provide predictive information so long as they precede the auditory stimulus and provide reliable cues (see Nahorna et al., 2012).

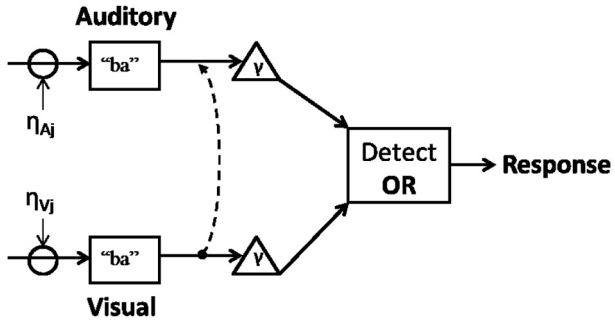


FIGURE 1 | Inputs interact with noise while evidence for a category (e.g., "ba") accumulates toward threshold (γ). Once enough information in either modality reaches threshold, a decision is made (e.g., "ba" vs. "da"). Visual information interacts with auditory cortical regions (dotted line) leading to updated priors. This model does not rule out the possibility that auditory cues can reciprocally influence viseme recognition.

A critical issue pertaining to visual predictive coding, then, relates to the "rigidity" of the internal rules (fixed priors). van Wassenhove (2013) discussed research suggesting the stability of priors/representations that are innate or otherwise become firmly established during critical developmental periods (Rosenblum et al., 1997; Lewkowicz, 2000). Lewkowicz (2000) argued that the ability to detect multisensory synchrony and match "duration and rate" are established early in life. In the domain of speech, Rosenblum and colleagues have argued that infants are sensitive to the McGurk effect and also to matched vs. mismatched articulatory movements and speech sounds.

While these studies suggest some rigidity of priors, I would emphasize that prior probabilities or "internal rules" remain malleable into adulthood. This adaptive perspective finds support among Bayesian theorists who argue that priors are continually updated in light of new evidence. Research indicates that differences in the ability to detect subtle auditory-visual asynchronies changes even into early adulthood (Hillock et al., 2011). Additionally, perceptual learning and adaptation techniques can alter priors in such a way that perceptions of asynchronies are modified via practice (Fujisaki et al., 2004; Vatakis et al., 2007; Powers et al., 2009) or experience with a second language (Navarra et al., 2010). Importantly, continual updating of "fixed" priors allows adult perceivers to (re)learn, fine tune, and adapt to multimodal signals across listening conditions, variable talkers, and attentional loads. van Wassenhove (2013) discussed how subjects can "automatically" match pitch and spatial frequency patterns (Evans and Treisman, 2010). This certainly shows that subjects can match auditory and visual information based on prior experience. Altieri et al. (2013) have also shown that adults can learn to match auditory and visual patterns more efficiently after only one day of practice! Reaction times and EEG signals indicated rapid learning and higher integration efficiency after only 1 h of training, followed by a period of gradual learning that remained stable over 1 week.

Such findings appear consistent with a unified parallel framework where visual information influences auditory processing and where visual predictability can be reweighted through learning. **Figure 1** represents an attempt to couch predictive coding within adaptive parallel accounts of integration.

ACKNOWLEDGMENTS

The research 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).

REFERENCES

Altieri, N., Pisoni, D. B., and Townsend, J. T. (2011). Behavioral, clinical, and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. *Seeing Perceiving* 24, 513–539. doi: 10.1163/187847611X595864

Altieri, N., Stevenson, R. A., Wallace, M. T., and Wenger, M. J. (2013). 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

Conrey, B., and Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *J. Acoust. Soc. Am.* 119, 4065. doi: 10.1121/1.2195091

Evans, K. K., and Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *J. Vis.* 10:6. doi: 10.1167/ 10.1.6

Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778. doi: 10.1038/nn1268

Hillock, A. R., Powers, A. R., and Wallace, M. T. (2011). Binding of sights and sounds: age-related changes in audiovisual temporal processing. *Neuropsychologia* 49, 461–467. doi: 10.1016/j.neuropsychologia.2010.11.041

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epignetic systems/limitations view. *Psychol. Bull.* 162, 281–308. doi: 10.1037/0033-2909.126.2.281

Massaro, D. W. (1987). "Speech perception by ear and eye," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum), 53–83.

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

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

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., and Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Res.* 1323, 84–93. doi: 10.1016/j.brainres.2010.01.059

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

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

Rosenblum, L., Schmuckler, M. A., and Johnson, J. A. (1997). The McGurk effect in infants. *Percept. Psychophys.* 59, 347–357. doi: 10.3758/BF03211902

Schroeder, C., 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

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

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

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 *The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (Hillsdale, NJ: LEA), 3–50.

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. (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

Vatakis, A., Navarra, J., Soto-Faraco, S., and Spence, C. (2007). Temporal recalibration during asynchronous audiovisual speech perception. *Exp. Brain Res.* 181, 173–181. doi: 10.1007/s00221-007-0918-z

Received: 12 November 2013; accepted: 10 March 2014; published online: 24 March 2014.

*Citation: Altieri N (2014) Multisensory integration, learning, and the predictive coding hypothesis. Front. Psychol. **5**:257. doi: 10.3389/fpsyg.2014.00257*

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

frontiers in PSYCHOLOGY

GENERAL COMMENTARY
published: 01 May 2014
doi: 10.3389/fpsyg.2014.00352

The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech

Ryan A. Stevenson[1]*, Mark T. Wallace[2,3,4,5,6] and Nicholas Altieri[7]

[1] *Psychology Department, University of Toronto, Toronto, ON, Canada*
[2] *Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, TN, USA*
[3] *Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN, USA*
[4] *Vanderbilt Kennedy Center, Vanderbilt University Medical Center, Nashville, TN, USA*
[5] *Department of Psychology, Vanderbilt University, Nashville, TN, USA*
[6] *Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, USA*
[7] *Department of Communication Sciences and Disorders, Idaho State University, Pocatello, ID, USA*
**Correspondence: ryan.andrew.stevenson@gmail.com*

Edited by:
Bruce D. McCandliss, Vanderbilt University, USA

Reviewed by:
Urs Maurer, University of Zurich, Switzerland

Keywords: multisensory processing, audiovisual integration, speech perception, temporal processing, sensory processing, crossmodal, perceptual binding, speech integration

A commentary on

Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs
by Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., and van Atteveldt, N. (2013). Front. Psychol. 4:331. doi: 10.3389/fpsyg.2013.00331

The amount of research focused on multisensory speech perception has expanded considerably in recent years. Much of this research has focused on which factors influence whether or not an auditory and a visual speech input are "integrated" (i.e., perceptually bound); a special case of how our perceptual systems solve the "binding problem" (Treisman, 1996). The factors that have been identified as influencing multisensory integration can be roughly divided into two groups. First are the low-level stimulus factors that include the physical characteristics of the sensory signals. The most commonly studied of these include the spatial (e.g., Macaluso et al., 2004; Wallace et al., 2004) and temporal (e.g., Miller and D'Esposito, 2005; Stevenson et al., 2011) relationship of the two inputs, and their relative effectiveness (e.g., James et al., 2012; Kim et al., 2012) in driving a neural, perceptual, or behavioral response. The second group of factors can be considered more higher-order or cognitive, and include factors such as the semantic congruence of the auditory and visual signals (Laurienti et al., 2004) or whether or not the gender of the speaker's voice is matched to the face (Lachs and Pisoni, 2004).

While these two categories can be considered conceptually distinct, they are related because of their mutual dependence upon the natural statistics of signals in the environment. When auditory and visual speech signals are closely proximate in time (low-level), they are more likely to have originated from the same speaker, and thus should be integrated (Dixon and Spitz, 1980; Stevenson et al., 2012b). Likewise, if an auditory and a visual speech signal are semantically congruent (high-level), they are more likely to have originated from the same speaker and thus should be integrated (Calvert et al., 2000). Given that these low- and high-level factors are each reflective of the natural statistics of the environmental signals, they will generally co-vary. Taking speech as an example, in a natural setting, the temporally-coincident auditory and visual components of a syllable or word are also semantically congruent (Spence, 2007).

To date, most research has investigated these low- and high-level factors independently. These studies have been highly informative, providing descriptions as to how each of these factors contributes to the process of multisensory integration. What has not received a great deal of focus is the interplay between these factors. A handful of experiments have investigated how low-level factors interact with one another and influence multisensory integration (Macaluso et al., 2004; Royal et al., 2009; Stevenson et al., 2012a), but few have attempted to bridge between low-level stimulus-characteristics and high-level cognitive factors (Vatakis and Spence, 2007). A recent article by Ten Oever et al. (2013), *Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs* addresses this gap in our understanding by investigating the interaction between stimulus timing and semantic congruency modulated by changes in place of articulation or voicing.

In this study, participants were presented with single-syllable stimuli, with auditory, visual, and audiovisual syllables systematically manipulated according to place of articulation and voicing. In addition, the temporal alignment of the audiovisual presentations was also parametrically varied. Hence, semantic content was varied through changes both in the auditory (voicing) and visual (place

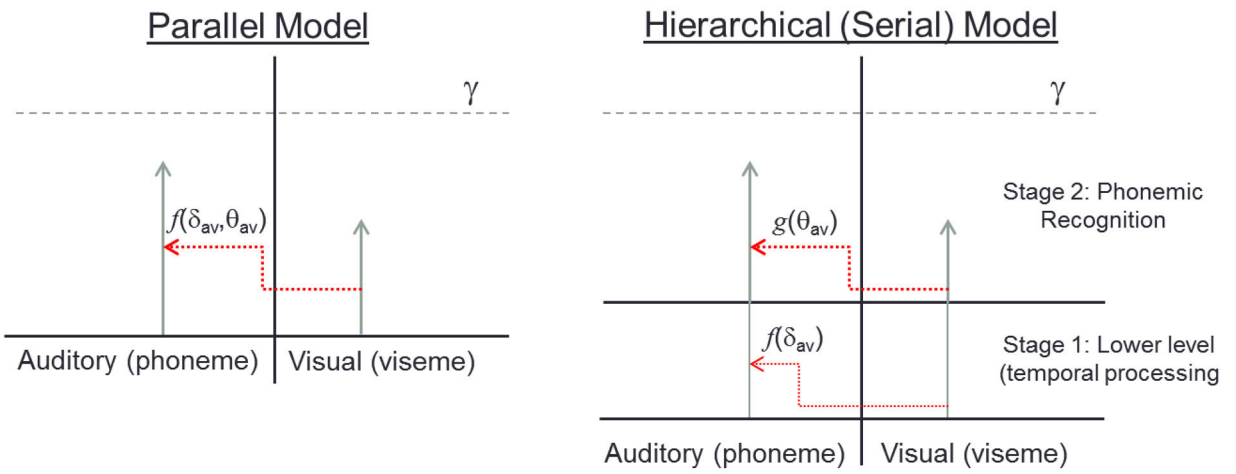


FIGURE 1 | The left panel shows a "parallel accumulator" model with auditory and visual evidence racing toward threshold (γ). The amount of visual influence on the auditory signal is a function, *f*, of parameters δ, and θ which represent temporal coincidence detection and phonetic congruence respectively, both of which contribute evidence to a single accumulator. The serial model on the right shows two separate stages where integration is affected first by temporal, then by semantic processing. Hence, in stage 1, visual information influences auditory processing only as a function *f* of temporal coincidence. In stage 2, visual information influences auditory processing solely as a function,*g*, of phonemic compatibility.

of articulation) signals, while at the same time, the relative timing of the auditory and visual stimuli were systematically varied. While the results specific to these factors are interesting on their own, most germane to this commentary is how these two factors interacted. The authors measured the window of time within which the visual cue influenced the syllable that was heard. This probabilistic construct, referred to as the "time window of integration" or the "temporal binding window," has been shown to vary greatly according the type of stimulus being integrated (Vatakis and Spence, 2006; Stevenson and Wallace, 2013). In the Ten Oever et al. study, semantically congruent stimuli were found to be associated with a wider temporal binding window than semantically incongruent stimuli. That is, stimulus components that are semantically matched have higher rates of integration at more temporally disparate offsets.

The result is surprising in that it runs counter to predictions generated by hierarchical serial models. In such models, lower-level properties such as stimulus timing are processed initially, and are then followed by the processing of the linguistic (i.e., semantic) content in the auditory and visual signals. However, the current results, by illustrating an interaction between timing and congruency, suggest that hierarchical models are insufficient to explain the data. Rather, we posit that these results are better interpreted within a "parallel accumulation of evidence" framework (**Figure 1**). In this model, the temporal relationship of two sensory inputs provides important information about the likelihood that those two inputs originated from the same speaker and should be integrated. In addition, the semantic congruence of these inputs also provides information as to whether or not the two sensory inputs should be bound. Importantly, these two types of evidence are pooled into a single decision criterion. Thus, within such a framework, when stimuli are semantically congruent, a decreased amount of temporal alignment is needed in order to cross a decision bound that would result in these two inputs being integrated, manifesting in a broader temporal binding window for semantically congruent speech stimulus pairs.

Through this interaction between stimulus timing and semantic congruence, Ten Oever and colleagues provided compelling evidence that low-level stimulus and high-level cognitive factors are not processed in a completely serial manner, but rather interact with one another in the formation of a perceptual decision. These results have significant implications in informing our view as to the neurobiological substrates involved in real-world multisensory perceptual processes. Most importantly, the work suggests that significant feedforward and feedback circuits are engaged in the processing of naturalistic multisensory stimuli, and that these circuits work in a parallel and cooperative fashion in evaluating the statistical relations of the stimuli to one another on both their low-level (i.e., stimulus feature) and high-level (i.e., learned semantic) correspondences.

REFERENCES

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

Dixon, N. F., and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception* 9, 719–721. doi: 10.1068/p090719

James, T. W., Stevenson, R. A., and Kim, S. (2012). "Inverse effectiveness in multisensory processing," in *The New Handbook of Multisensory Processes*, ed B. E. Stein (Cambridge, MA: MIT Press), 207–221.

Kim, S., Stevenson, R. A., and James, T. W. (2012). Visuo-haptic neuronal convergence demonstrated with an inversely effective pattern of BOLD activation. *J. Cogn. Neurosci.* 24, 830–842. doi: 10.1162/jocn_a_00176

Lachs, L., and Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecol. Psychol.* 16, 159–187. doi: 10.1207/s15326969eco1603_1

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., and Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Exp. Brain Res.* 158, 405–414. doi: 10.1007/s00221-004-1913-2

Macaluso, E., George, N., Dolan, R., Spence, C., and Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21, 725–732. doi: 10.1016/j.neuroimage.2003.09.049

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

Royal, D. W., Carriere, B. N., and Wallace, M. T. (2009). Spatiotemporal architecture of cortical receptive fields and its impact on multisensory interactions. *Exp. Brain Res.* 198, 127–136. doi: 10.1007/s00221-009-1772-y

Spence, C. (2007). Audiovisual multisensory integration. *Acoust. Sci. Technol.* 28, 61–70. doi: 10.1250/ast.28.61

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., and Wallace, M. T. (2012a). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. *Exp. Brain Res.* 219.1, 121–137. doi: 10.1007/s00221-012-3072-1

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., and Wallace, M. T. (2013). Multisensory temporal integration: task and stimulus dependencies. *Exp. Brain Res.* 227, 249–261. doi: 10.1007/s00221-013-3507-3

Stevenson, R. A., Zemtsov, R. K., and Wallace, M. T. (2012b). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *J. Exp. Psychol. Hum. Percept. Perform*. 38.6, 1517. doi: 10.1037/a0027339

Ten Oever, S., Sack, A. T., 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

Treisman, A. (1996). The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178. doi: 10.1016/S0959-4388(96)80070-5

Vatakis, A., and Spence, C. (2006). 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. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys.* 69, 744–756. doi: 10.3758/BF03193776

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Exp. Brain Res.* 158, 252–258. doi: 10.1007/s00221-004-1899-9

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 January 2014; accepted: 03 April 2014; published online: 01 May 2014.

Citation: Stevenson RA, Wallace MT and Altieri N (2014) The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. Front. Psychol. ***5****:352. doi: 10.3389/fpsyg.2014.00352*

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

frontiers in PSYCHOLOGY

GENERAL COMMENTARY
published: 11 April 2014
doi: 10.3389/fpsyg.2014.00297

Caregiver influence on looking behavior and brain responses in prelinguistic development

Heather L. Ramsdell-Hudock*

Infant Vocal Development Laboratory, Communication Sciences and Disorders, Idaho State University, Pocatello, ID, USA
**Correspondence: ramsdell@isu.edu*

Edited by:
Nicholas Altieri, Idaho State University, USA

Reviewed by:
Julie Gros-Louis, University of Iowa, USA

Keywords: looking behavior, brain response, prelinguistic vocal development, caregiver-infant interaction, audiovisual speech integration

A commentary on

Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life
by Kushnerenko, E., Tomalski, P., Ballieux, H., Potton, A., Birtles, D., Frostick, C., et al. (2013). Front. Psychol. 4:432. doi: 10.3389/fpsyg.2013.00432

Much is known about the development of speech perception, but there remain gaps in understanding in the development of speech production. Social interaction is a popular topic among researchers of human development, and the development of speech production is clearly dependent upon social interaction (Goldstein et al., 2003; Goldstein and Schwade, 2008). The importance of social interaction becomes particularly apparent when infants begin to speak the language of their ambient environment (Lewkowicz and Hansen-Tift, 2012). Prior to first word production, however, an exchange between caregivers and infants occurs, such that infant vocalizations and other changes in development directly influence caregiver responses, which in turn influence various components of interaction.

LOOKING BEHAVIOR AND BRAIN RESPONSE

When presented with audiovisual stimuli of speaking faces between 6 and 12 months of age, researchers have compiled information on which parts of the face infants scan (e.g., around the eyes or the mouth), areas of brain activation, and associations between looking behavior and brain responses with later expressive and receptive language abilities. For looking behavior, a developmental shift in attention to audiovisual speech has been demonstrated between 4 and 8 months of age, with an increase in the time spent looking at the mouth as compared to eyes (Lewkowicz and Hansen-Tift, 2012). Increased looking behavior toward the mouth may support understanding of language, as faces convey an abundance of linguistic information (Wagner et al., 2013). Longer time looking to the eyes may indicate searching for additional social cues to resolve ambiguity in audiovisual stimuli (Kushnerenko et al., 2013).

For brain responses, significant negative correlations have been observed between auditory comprehension scores and the amplitude of the infantile P2 over right frontocentral brain regions in response to ambiguous audiovisual speech stimuli. The correlation weakens during the first year of development, presumably as auditory recognition improves. Infants with less precise, or less developed auditory speech processing at 6–9 months of age may rely more heavily on visual cues when ambiguous speech stimuli are presented (Kushnerenko et al., 2013). An alternative hypothesis is that infants may begin to attend more to the mouth region of a communication partner and process audiovisual information outside of the right lateralized frontocentral brain regions (Kushnerenko et al., 2013).

PRELINGUISTIC VOCAL DEVELOPMENT AND CAREGIVER-INFANT INTERACTION

The shift in visual attention may also be related to vocal development. In particular, developmental advances occurring with respect to speech production between 6 and 12 months of age directly influence caregiver interaction. Caregivers begin to attend more, and respond differently to well-formed infant productions. This in turn, may lead infants to allocate visual attention to different regions of the face, and hence, encourage the eyes-to-mouth-to-eyes attentional shift.

Infants normally transition through several stages of vocal development in the first year of life. In the phonation stage, from birth to 2 months of age, infants gradually become more able to manipulate normal phonation in production of quasivowels. In the primitive articulation stage, from 1 to 4 months of age, infants gradually become more capable of manipulating their vocal tract during voicing in production of "cooing" and "gooing" sounds. In the expansion stage, from 3 to 8 months of age, infants gradually become more able to open and posture their vocal tract in production of full vowels and marginal babbles (Oller, 2000). During these first months of life, caregivers are likely to simply gauge infant well-being from phonation, and to imitate vocalizations produced by the infant (Julien and Munson, 2012; Olson and Masur, 2012). This sort of action from the caregiver does not enhance vocalizations and may not draw the infant's attention toward the mouth—the visible part of the speech mechanism.

In the canonical stage, from approximately 5–10 months of age, infants gradually become more able to make well-timed movements of their articulators from open to closed postures in production of canonical babbling (Oller, 2000). In this final stage, prior to production of

words, infants begin to produce syllables, potential components of words (Oller, 1980; Koopmans-van Beinum and van der Stelt, 1986; Stark et al., 1993). When caregivers identify these babbled syllables, they intuitively begin to interact with the infant around them, treating them as potential words (Veneziano, 1988; Stoel-Gammon, 2011). As word learning begins, caregivers engage infants, recognizing babbled syllables and their potential relation with the ambient language (Ramsdell et al., 2012; Oller et al., 2013). In this stage, infant vocalizations are well-formed and more familiar, and caregivers begin to encourage language growth through expanding on and enhancing these productions (Gros-Louis et al., 2006; Olson and Masur, 2012). Simultaneously, infants are learning from caregiver input. In a study conducted by Goldstein and Schwade, sixty 9.5 month old infants were found to begin producing more speech-like forms in coordination with caregiver response to their vocalizations (2008). At this point in development, caregiver response to infant vocalizations may draw the infant's attention toward the mouth, and the infant then receives multimodal (auditory and visual) feedback supporting productions, which in turn help to facilitate vocal development.

Still, there are other interpretations that may also contribute to shifts in attention. Perhaps not only changing caregiver/infant interaction, but also changing social-cognitive abilities for joint attention later in the first year of development, shift attention from the mouth, back to the eyes. Infants engage in more joint attention episodes between 9 and 12 months of age, with improving skill for triadic interactions from people to objects (Bakeman and Adamson, 1984). Further, infants are utilizing new forms of nonverbal communication during interactions, such as pointing, showing, and giving (Bates et al., 1975). During this time, joint attention and gesture use occur together with coordinating eye gaze between objects and social partners (Messinger and Fogel, 1998; Wu and Gros-Louis, 2014).

CONCLUSION

As prelinguistic vocalizations develop and become well-formed, caregiver responses change to *teach* language. This changing caregiver-infant interaction could help to guide the infant in attending to different areas of the face, thereby causing processing to shift to different neural circuitry. Accordingly, the changes occurring in looking behavior and brain responses are likely not to be solely dependent upon the infants own endogenous attentional mechanisms and motivations to vocalize, but dependent on caregiver interaction as well.

REFERENCES

Bakeman, R., and Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Dev.* 55, 1278–1289. doi: 10.2307/1129997

Bates, E., Luigia, C., and Volterra, V. (1975). The acquisition of performatives prior to speech. *Merrill Palmer Q. Behav. Dev.* 21, 205–266.

Goldstein, M. H., King, A. P., and West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8030–8035. doi: 10.1073/pnas.1332441100

Goldstein, M. H., and Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychol. Sci.* 19, 515–523. doi: 10.1111/j.1467-9280.2008.02117.x

Gros-Louis, J., West, M. J., Goldstein, M. H., and King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *Int. J. Behav. Dev.* 30, 112–119. doi: 10.1177/0165025406071914

Julien, H. M., and Munson, B. (2012). Modifying speech to children based on their perceived phonetic accuracy. *J. Speech Lang. Hear. Res.* 55, 1836–1848. doi: 10.1044/1092-4388(2012/11-0131)

Koopmans-van Beinum, F. J., and van der Stelt, J. M. (1986). "Early stages in the development of speech movements," in *Precursors of Early Speech*, eds B. Lindblom and R. Zetterstrom (New York, NY: Stockton Press), 37–50.

Kushnerenko, E., Tomalski, P., Ballieux, H., Potton, A., Birtles, D., Frostick, C., et al. (2013). Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life. *Front. Psychol.* 4:432. doi: 10.3389/fpsyg.2013.00432

Lewkowicz, D. J., and Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1431–1436. doi: 10.1073/pnas.1114783109

Messinger, D. S., and Fogel, A. (1998). Give and take: the development of conventional gestures. *Merrill Palmer Q.* 44, 566–590.

Oller, D. K. (1980). "The emergence of the sounds of speech in infancy," in *Child Phonology, Vol 1: Production*, eds G. Yeni-Komshian, J. Kavanagh, and C. Ferguson (New York, NY: Academic Press), 93–112.

Oller, D. K. (2000). *The Emergence of the Speech Capacity.* Mahwah, NJ: Laurence Erlbaum.

Oller, D. K., Buder, E. H., Warlaumont, A., Ramsdell-Hudock, H. L., Iyer, S. N., Franklin, B., et al. (2013). "Infant vocal development: the search for early identification of disorders," in *Seminar Presented at the American Speech-Language-Hearing Association Annual Convention* (Chicago, IL).

Olson, J., and Masur, E. F. (2012). Mothers respond differently to infants' familiar versus non-familiar verbal imitations. *J. Child Lang.* 39, 731–752. doi: 10.1017/S0305000911000262

Ramsdell, H. L., Oller, D. K., Buder, E. H., Ethington, C. A., and Chorna, L. (2012). Identification of prelinguistic phonological categories. *J. Speech Lang. Hear. Res.* 55, 1626–1639. doi: 10.1044/1092-4388(2012/11-0250)

Stark, R. E., Bernstein, L. E., and Demorest, M. E. (1993). Vocal communication in the first 18 months of life. *J. Speech Hear. Res.* 36, 548–558.

Stoel-Gammon, C. (2011). Relationships between lexical and phonological development in young children. *J. Child Lang.* 38, 1–34. doi: 10.1017/S0305000910000425

Veneziano, E. (1988). "Vocal-verbal interaction and the construction of early lexical knowledge," in *The Emergent Lexicon: The Child's Development of a Linguistic Vocabulary*, eds M. D. Smith and J. L. Locke (San Diego, CA: Academic Press), 109–147.

Wagner, J. B., Luyster, R. J., Yim, J. Y., Tager-Flusberg, H., and Nelson, C. A. (2013). The role of early visual attention in social development. *Int. J. Behav. Dev.* 37, 118–124. doi: 10.1177/0165025412468064

Wu, Z., and Gros-Louis, J. (2014). Infants' prelinguistic communicative acts and maternal responses: relations to linguistic development. *First Lang.* 34, 72–90. doi: 10.1177/0142723714521925

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 January 2014; accepted: 21 March 2014; published online: 11 April 2014.

*Citation: Ramsdell-Hudock HL (2014) Caregiver influence on looking behavior and brain responses in prelinguistic development. Front. Psychol. **5**:297. doi: 10.3389/fpsyg.2014.00297*

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