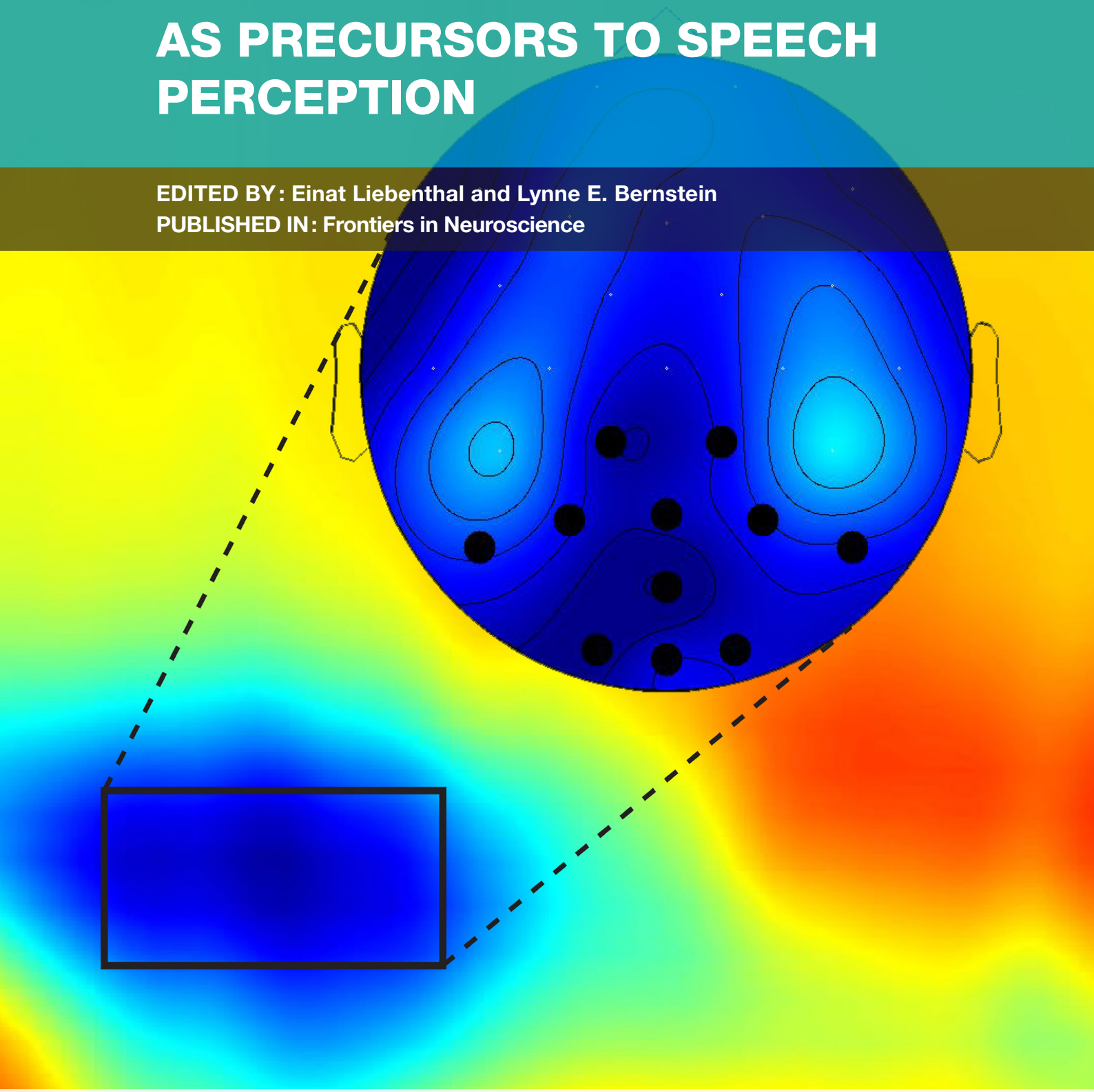
NEURAL MECHANISMS OF PERCEPTUAL CATEGORIZATION AS PRECURSORS TO SPEECH PERCEPTION

EDITED BY: Einat Liebenthal and Lynne E. Bernstein
PUBLISHED IN: Frontiers in Neuroscience



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-158-6
DOI 10.3389/978-2-88945-158-6

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.

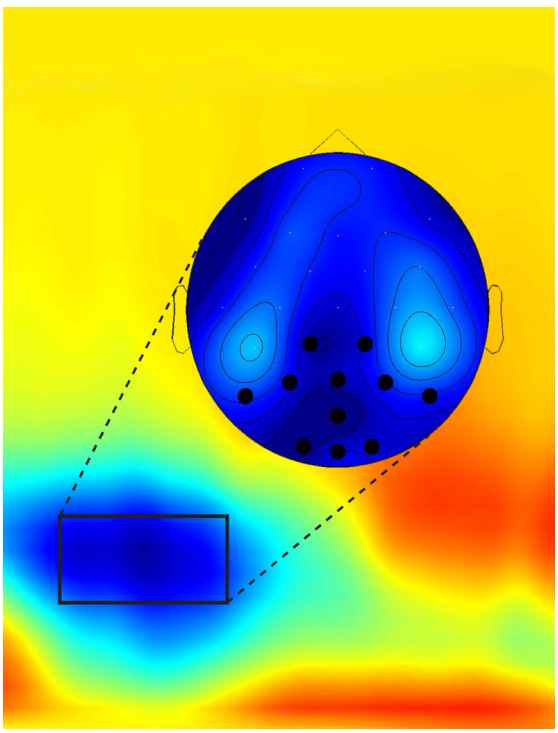
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

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**

NEURAL MECHANISMS OF PERCEPTUAL CATEGORIZATION AS PRECURSORS TO SPEECH PERCEPTION

Topic Editors:
Einat Liebenthal, Harvard Medical School, USA
Lynne E. Bernstein, George Washington University, USA



Averaged time-frequency representation of EEG for a tone categorization condition. Warmer colors indicate increases in power (compared to baseline), cooler colors indicate decreases (suppressions). The strongest suppression effect is found in the alpha range between 7 and 11 Hz and in a time window between 400–700 ms after tone onset. The inlet shows the topography of this effect, with maxima at central-posterior electrodes (selection marked with black dots). From Scharinger et al., in this ebook.

Perceptual categorization is fundamental to the brain's remarkable ability to process large amounts of sensory information and efficiently recognize objects including speech. Perceptual categorization is the neural bridge between lower-level sensory and higher-level language processing.

A long line of research on the physical properties of the speech signal as determined by the anatomy and physiology of the speech production apparatus has led to descriptions of the acoustic information that is used in speech recognition (e.g., stop consonants place and manner of articulation, voice onset time, aspiration). Recent research has also considered what visual cues are relevant to visual speech recognition (i.e., the visual counter-parts used in lipreading or audiovisual speech perception).

Much of the theoretical work on speech perception was done in the twentieth century without the benefit of neuroimaging technologies and models of neural representation. Recent progress in understanding the functional organization of sensory and association cortices based on advances in neuroimaging presents the possibility of achieving a comprehensive and far reaching account of perception in the service of language. At the level of cell assemblies, research in animals and humans suggests that neurons in the temporal cortex are important for encoding biological categories. On the cellular level, different classes of neurons (interneurons and pyramidal neurons) have been suggested to play differential roles in the neural computations underlying auditory and visual categorization.

The moment is ripe for a research topic focused on neural mechanisms mediating the emergence of speech representations (including auditory, visual and even somatosensory based forms). Important progress can be achieved by juxtaposing within the same research topic the knowledge that currently exists, the identified lacunae, and the theories that can support future investigations. This research topic provides a snapshot and platform for discussion of current understanding of neural mechanisms underlying the formation of perceptual categories and their relationship to language from a multidisciplinary and multisensory perspective. It includes contributions (reviews, original research, methodological developments) pertaining to the neural substrates, dynamics, and mechanisms underlying perceptual categorization and their interaction with neural processes governing speech perception.

Citation: Liebenthal, E., Bernstein, L. E., eds. (2017). Neural Mechanisms of Perceptual Categorization as Precursors to Speech Perception. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-158-6

Table of Contents

frontiers in Neuroscience

EDITORIAL
published: 14 February 2017
doi: 10.3389/fnins.2017.00069

Editorial: Neural Mechanisms of Perceptual Categorization as Precursors to Speech Perception

Einat Liebenthal[1]* *and Lynne E. Bernstein*[2]

[1] Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [2] Department of Speech and Hearing Science, George Washington University, Washington, DC, USA

Keywords: speech perception, phonemic perception, categorization, category learning, auditory processing, audiovisual processing, neural mechanism, neuroimaging

Editorial on the Research Topic

Neural Mechanisms of Perceptual Categorization as Precursors to Speech Perception

This research topic describes recent advances in understanding the brain functional organization for sensory categorization along with its implications for speech perception. Among the 14 papers, one theme is how neural representations of auditory and visual input are transformed across different scales of neural organization to enable speech perception, and another is the neural mechanisms of category learning.

In the first theme, several animal and human studies delve into the complex hierarchical organization of auditory ventral pathways for speech perception. Prior work has established an important role for the auditory ventral stream in complex sound categorization (Rauschecker and Scott, 2009; Romanski and Averbeck, 2009). In humans, a preference has convincingly been demonstrated for phonemic over non-phonemic sounds in non-primary auditory fields in the middle of the ventrolateral superior temporal cortex (mSTG/S) (Liebenthal et al., 2005; Leaver and Rauschecker, 2010). The present papers contribute novel insights about the function of dorsal areas in and near the auditory core, the functional specificity of the mSTG/S, and the role of non-auditory areas, for phonemic perception. Collectively, they suggest that multiple stages of abstraction from the original form of speech occur in low-level sensory cortices. In the mSTG/S, the neural representations are highly specific to phonemic categories.

Tsunada and Cohen's review of research in the monkey suggests that single neurons in the auditory core encode categories for simple sounds (e.g., direction of spectral changes), whereas neurons in the auditory belt encode more complex categories (including speech phonemes) based on input from the entire population of core neurons. At the cellular level, they report the intriguing finding that different classes of neurons within the auditory belt may have different sensitivity to category information: The more common pyramidal neurons encode auditory categories with less sensitivity than the less common interneurons (Tsunada et al., 2012). Astikainen et al. also show that in anesthetized rats' primary auditory cortex, neurons automatically encode structural patterns (order of syllable repetition) from a fast paced speech stream and generalize to novel patterns.

Based on intracranial high-gamma electrophysiological recordings in subjects with intractable epilepsy, Steinschneider et al. propose that within 200 ms, activity in the human primary and non-primary auditory cortices reflects non-categorical spectrotemporal sound attributes. Only later, activity in non-primary auditory areas receiving modulatory input from higher-order, lexico-semantic associative cortex represents phoneme categories.

Edited by:
Isabelle Peretz,
Université de Montréal, Canada

Reviewed by:
Ingrid Johnsrude,
University of Western Ontario, Canada

***Correspondence:**
Einat Liebenthal
eliebenthal@partners.org

Specialty section:
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *26 October 2016*
Accepted: *31 January 2017*
Published: *14 February 2017*

Citation:
Liebenthal E and Bernstein LE (2017) Editorial: Neural Mechanisms of Perceptual Categorization as Precursors to Speech Perception. Front. Neurosci. 11:69. doi: 10.3389/fnins.2017.00069

Using multivariate pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data, Joanisse and Desouza suggest that primary and non-primary areas in the human auditory cortex encode the direction of frequency modulations of complex non-speech sounds. Using an fMRI adaptation paradigm, Humphries et al. show that a relatively large area in the dorsolateral superior temporal cortex is sensitive to complex acoustic patterns in phonemic and non-phonemic sounds, whereas a small portion of the ventrolateral superior temporal cortex responds specifically to phonemic sounds, with relatively little overlap between the areas. In addition, an area of the medial superior temporal plane shows a preference for non-phonemic sounds. The results support a multi-stage hierarchical stream for speech perception extending from the superior temporal plane to the superior temporal sulcus.

Liebenthal et al. present a large meta-analysis of neuroimaging studies of the left superior temporal cortex, and find a strong preference for speech perception over other language functions in the mSTG/S. This area preferred linguistic over non-linguistic input and auditory over visual processing, prompting the suggestion that a high functional specificity of the left mSTS for auditory speech may be an important means by which the human brain achieves its exquisite affinity and efficiency for native speech perception.

Bernstein and Liebenthal's review of visual speech proposes a neural model of speech perception according to which visual aspects of speech are represented hierarchically in ascending visual pathways, with a functional organization similar to that of auditory pathways. Central to the model is the proposal that a visual area in the left posterior temporal cortex represents visual phoneme categories.

The second theme concerns how altered experience and training regimes affect perceptual categorization and neural processes. Current understanding of the normative organization of speech categories is based mostly on experiments with adults who have experienced normal language acquisition and who listen in their native language. Experiments that use natural or artificial factors that perturb and change the system help to further define the organization and mechanisms of categorization.

Heald et al. report on pitch categorization. They suggest that individuals vary in the extent to which they rely on an internal systematic tone organization. Absolute pitch (AP) possessors may be more analogous to speech perceivers than non-AP musical experts, and musical novices are expected to be least able to categorize tones based on internal organization. All three types of participants were influenced by the structure of the stimulus set and possessed useful prior pitch knowledge. Increased expertise was associated with greater influence of internal category structure.

Myers reviews the literature on normative category processing and suggests that second-language learning involves remapping the native language perceptual space to the perceptual space of the second language. Training studies typically use explicit category training, and Myers points to a wide network of frontal and temporal areas that is recruited as a result of such training. She suggests that learned sensitivity to categories is first observable in the frontal lobe and with greater expertise is observable in temporal areas. This shift is consistent with the reverse hierarchy theory (Ahissar and Hochstein, 1997; Ahissar et al., 2008) and with frontal-to-temporal feedback as a mechanism that assists in warping category representations for the second language.

Callan et al. report an fMRI study comparing English and Japanese speakers listening to native and accented English /r/-/l/. The accented English of Japanese natives is difficult for native English speakers and the English /r/-/l/ is a difficult distinction for native Japanese speakers. In their results, temporal cortex areas are not significantly modulated by expertise. Instead, more difficult distinctions recruit the right cerebellum and left premotor cortex (PMC) in both groups. Second language listening additionally recruits the right PMC and left cerebellum.

Ley et al. discuss the value of MVPA for revealing high plasticity of sound representation in auditory temporal areas as a function of experience and learning. They suggest that sensory plasticity and attention processes interact to mediate category learning. They review findings within predictive coding models of perceptual learning and categorization that support a hierarchical architecture in which variation in sensory information confronts top-down signals that update bottom-up representations.

Scharinger et al. discuss the role of auditory attention in realistic listening conditions, when perception needs to adapt to dynamic degradation of certain stimulus cues. They use multimodal neuroimaging of oscillatory activity in the alpha band to study auditory categorization and highlight the role of posterior auditory areas and the inferior parietal cortex for optimal utilization of informative stimulus cues and inhibition of uninformative cues.

Lim et al. approach speech categorization through the perspective of cognitive neuroscience models that attempt to account for multiple learning systems and corresponding neural structures. These authors frame questions about the relationships between frontal and temporal cortices during learning within larger networks that include the basal ganglia. They discuss different types of feedback and task structure that may eventuate in different types of learning, declarative vs. procedural (Ashby et al., 1998). Category training tasks that encourage trainees to engage in *explicit* attempts to discover categorization rules or structure (declarative learning) result in limited generalization for speech categories, which are inherently multidimensional and incommensurate. Speech category learning appears to require procedural learning that involves bottom-up integration of stimulus features and dopaminergic reward signals.

Bernstein et al. behavioral study demonstrates an advantage to training with audiovisual speech in order to obtain improvements in the auditory perception of vocoded speech. Training used a paired associates task for which participants attempted to learn the associations between disyllabic non-sense words and non-sense pictures. Feedback was for association choices and not the phonemic content of the training stimuli. The audiovisual advantage

is interpreted within a multisensory extension of reverse hierarchy theory (Ahissar and Hochstein, 1997; Ahissar et al., 2008): Higher-level visual speech representations during audiovisual training may guide the top-down search for to-be-learned acoustic phonetic features. The training task may also promote procedural learning of the type described by Lim et al.

Future research should build on these insights to advance understanding of the neural basis of speech perception and learning.

AUTHOR CONTRIBUTIONS

EL and LB contributed equally to the conceptualization and editing of the research topic. EL and LB wrote the editorial together.

FUNDING

The work was supported by NIH R01 DC006287 and NIH R21 DC012634.

REFERENCES

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

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

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* 105, 442–481. doi: 10.1037/0033-295X.105.3.442

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

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

Romanski, L. M., and Averbeck, B. B. (2009). The primate cortical auditory system and neural representation of conspecific vocalizations. *Annu. Rev. Neurosci.* 32, 315–346. doi: 10.1146/annurev.neuro.051508.135431

Tsunada, J., Lee, J. H., and Cohen, Y. E. (2012). Differential representation of auditory categories between cell classes in primate auditory cortex. *J. Physiol.* 590, 3129–3139. doi: 10.1113/jphysiol.2012.232892

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

frontiers in
NEUROSCIENCE

REVIEW ARTICLE
published: 17 June 2014
doi: 10.3389/fnins.2014.00161

Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits

***Joji Tsunada*[1]* *and Yale E. Cohen*[1,2,3]**

[1] *Department of Otorhinolaryngology-Head and Neck Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[2] *Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA*
[3] *Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA*

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Amy Poremba, University of Iowa, USA
Rutvik Desai, University of South Carolina, USA

****Correspondence:***
Joji Tsunada, Department of Otorhinolaryngology-Head and Neck Surgery, Perelman School of Medicine, University of Pennsylvania, 3400 Spruce Street, 5 Ravdin, Philadelphia, PA 19104, USA
Office: Perelman School of Medicine, University of Pennsylvania, John Morgan Building B50, 3620 Hamilton Walk, Philadelphia, PA 19104-6055, USA
e-mail: tsunada@mail.med.upenn.edu

Categorization enables listeners to efficiently encode and respond to auditory stimuli. Behavioral evidence for auditory categorization has been well documented across a broad range of human and non-human animal species. Moreover, neural correlates of auditory categorization have been documented in a variety of different brain regions in the ventral auditory pathway, which is thought to underlie auditory-object processing and auditory perception. Here, we review and discuss how neural representations of auditory categories are transformed across different scales of neural organization in the ventral auditory pathway: from across different brain areas to within local microcircuits. We propose different neural transformations across different scales of neural organization in auditory categorization. Along the ascending auditory system in the ventral pathway, there is a progression in the encoding of categories from simple acoustic categories to categories for abstract information. On the other hand, in local microcircuits, different classes of neurons differentially compute categorical information.

Keywords: auditory category, ventral auditory pathway, speech sound, vocalization, pyramidal neuron, interneuron

INTRODUCTION

Auditory categorization is a computational process in which sounds are classified and grouped based on their acoustic features and other types of information (e.g., semantic knowledge about the sounds). For example, when we hear the word "Hello" from different speakers, we can categorize the gender of each speaker based on the pitch of the speaker's voice. On the other hand, in order to analyze the linguistic content transmitted by speech sounds, we can ignore the unique pitch, timbre etc. of each speaker and categorize the sound into the distinct word category "Hello." Thus, auditory categorization enables humans and non-human animals to extract, manipulate, and efficiently respond to sounds (Miller et al., 2002, 2003; Russ et al., 2007; Freedman and Miller, 2008; Miller and Cohen, 2010).

A specific type of categorization is called "categorical perception" (Liberman et al., 1967; Kuhl and Miller, 1975, 1978; Kuhl and Padden, 1982, 1983; Kluender et al., 1987; Pastore et al., 1990; Lotto et al., 1997; Sinnott and Brown, 1997; Holt and Lotto, 2010). The primary characteristic of categorical perception is that the perception of a sound does not smoothly vary with changes in its acoustic features. That is, in certain situations, small changes in the physical properties of an acoustic stimulus can cause large changes in a listener's perception of a sound. In other situations, large changes can cause no change in perception. The stimuli, which cause these large changes in perception, straddle the boundary between categories. For example, when we hear a continuum of smoothly varying speech sounds (i.e., a continuum of morphed stimuli between the phoneme prototypes "ba" and "da"), we experience a discrete change in perception. Specifically, a small change in the features of a sound near the middle of this continuum (i.e., at the category boundary between a listener's perception of "ba" and "da") will cause a large change in a listener's perceptual report. In contrast, when that same small change occurs at one of the ends of the continuum, there is little effect on the listener's report.

Even though some perceptual categories have sharp boundaries, the locations of the boundary are somewhat malleable. For instance, the perception of a phoneme can be influenced by the phonemes that come before it. When morphed stimuli, which are made from the prototypes "da" and "ga," are preceded by presentations of "al" or "ar," the perceptual boundary between the two prototypes shifts (Mann, 1980). Specifically, listeners' reports are biased toward reporting the morphed stimuli as "da" when it is preceded by "ar." When this morphed stimulus is preceded by "al," listeners are biased toward reporting the morphed stimulus as "ga."

Categories are not only formed based on the perceptual features of stimuli but also on more "abstract" types of information. An abstract category is one in which a group of arbitrary stimuli are linked together as a category based on some shared features, a common functional characteristic, semantic information,

or acquired knowledge. For instance, a combination of physical characteristics and knowledge about their reproductive processes puts dogs, cats, and killer whales into one category ("mammals"), but birds into a separate category. However, if we use different criteria to form a category of "pets," dogs, cats, and birds would be members of this "pet" category but not killer whales.

Behavioral responses to auditory communication signals (i.e., species-specific vocalizations) also provide evidence for abstract categorization. One example is the categorization of food-related species-specific vocalizations by rhesus monkeys (Hauser and Marler, 1993a,b; Hauser, 1998; Gifford et al., 2003). In rhesus monkeys, a vocalization called a "harmonic arch" transmits information about the discovery of rare, high-quality food. A different vocalization called a "warble" also transmits the same type of information: the discovery of rare, high-quality food. Importantly, whereas both harmonic arches and warbles transmit the same type of information, they have distinct spectrotemporal properties. Nevertheless, rhesus monkeys' responses to those vocalizations indicate that monkeys categorize these two calls based on their transmitted information and not their acoustic features. In another example, Diana monkeys form abstract-categorical representations for predator-specific alarm calls independent of the species generating the signal. Diana monkeys categorize and respond similarly to alarm calls that signify the presence of a leopard, regardless of whether the alarm calls are elicited from a Diana monkey or a crested guinea fowl (Zuberbuhler and Seyfarth, 1997; Züberbuhler, 2000a,b). Similarly, Diana monkeys show similar categorical-responses to eagle alarm calls that can be elicited from other Diana monkeys or from putty-nose monkeys (Eckardt and Zuberbuhler, 2004).

In order to better understand the mechanisms that underlie auditory categorization, it is essential to examine how neural representations of auditory categories are formed and transformed across different scales of neural organization: from across different brain areas to within local microcircuits. In this review, we discuss the representation of auditory categories in different cortical regions of the ventral auditory pathway; the hierarchical processing of categorical information along the ventral pathway; and the differential role that excitatory pyramidal neurons and inhibitory interneurons (i.e., different neuron classes) contribute to these categorical computations.

The ventral pathway is targeted because neural computations in this pathway are thought to underlie sound perception, which is critically related to auditory categorization and auditory scene analysis (Rauschecker and Scott, 2009; Romanski and Averbeck, 2009; Bizley and Cohen, 2013). The ventral auditory pathway begins in the core auditory cortex (in particular, the primary auditory cortex and the rostral field R) and continues into the anterolateral and middle-lateral belt regions. These belt regions then project either directly or indirectly to the ventral prefrontal cortex (**Figure 1**) (Hackett et al., 1998; Rauschecker, 1998; Kaas and Hackett, 1999, 2000; Kaas et al., 1999; Romanski et al., 1999a,b; Rauschecker and Tian, 2000; Rauschecker and Scott, 2009; Romanski and Averbeck, 2009; Recanzone and Cohen, 2010; Bizley and Cohen, 2013).

NEURAL TRANSFORMATIONS ACROSS CORTICAL AREAS IN THE VENTRAL AUDITORY PATHWAY

In this section, we discuss how auditory categories are processed in the ventral auditory pathway. More specifically, we review the representation of auditory categories across different regions in the ventral auditory pathway and then discuss the hierarchical processing of categorical information in the ventral auditory pathway.

Before we continue, it is important to define the concept of a "neural correlate of categorization." One simple definition is the following: a neural response is "categorical" when the responses are invariant to the stimuli that belong to the same category. In practice, neuroimaging techniques define "categorical" responses as equivalent activations of distinct brain regions by within-category stimuli and the equivalent activation of different brain regions by stimulus exemplars from a second category (Binder et al., 2000; Altmann et al., 2007; Doehrmann et al., 2008; Leaver and Rauschecker, 2010). At the level of single neurons, a neuron is said to be "categorical" if its firing rate is invariant to different members of one category and if it has a second level of (invariant) responsivity to stimulus exemplars from a second category (Freedman et al., 2001; Tsunada et al., 2011). The specific mechanisms that underlie the creation of category sensitive neurons are not known. However, presumably, they rely on the computations that mediate stimulus invariance in neural selectivity and perception (Logothetis and Sheinberg, 1996; Holt and Lotto, 2010; Dicarlo et al., 2012). Moreover, because animals can form a wide range of categories based on individual experiences, a degree of learning and plasticity must be involved in the creation of *de-novo* category selective responses (Freedman et al., 2001; Freedman and Assad, 2006). Indeed, when monkeys were trained to categorize stimuli with different category boundaries, boundaries for categorical responses in some brain areas (e.g., the prefrontal and parietal cortices) also changed (Freedman et al., 2001; Freedman and Assad, 2006).

HOW DO DIFFERENT CORTICAL AREAS IN THE VENTRAL AUDITORY PATHWAY SIMILARLY OR DIFFERENTIALLY REPRESENT CATEGORICAL INFORMATION?

It is well known that neurons become increasingly sensitive to more complex stimuli and abstract information between the beginning stages of the ventral auditory pathway (i.e., the core) and the latter stages (e.g., the ventral prefrontal cortex). For example, neurons in the core auditory cortex are more sharply tuned for tone bursts than neurons in the lateral belt (Rauschecker et al., 1995), whereas lateral-belt neurons are more sensitive to the spectrotemporal properties of complex sounds, such as vocalizations (Rauschecker et al., 1995; Tian and Rauschecker, 2004). Furthermore, beyond the auditory cortex, the ventral prefrontal cortex not only encodes complex sounds (Averbeck and Romanski, 2004; Cohen et al., 2007; Russ et al., 2008a; Miller and Cohen, 2010) but also has a critical role for attention and memory-related cognitive functions (e.g., memory retrieval) which are critical for abstract categorization (Goldman-Rakic, 1995; Miller, 2000; Miller and Cohen, 2001; Miller et al., 2002, 2003; Gold and Shadlen, 2007; Osada et al., 2008; Cohen et al., 2009; Plakke et al., 2013a,b,c; Poremba et al., 2013).

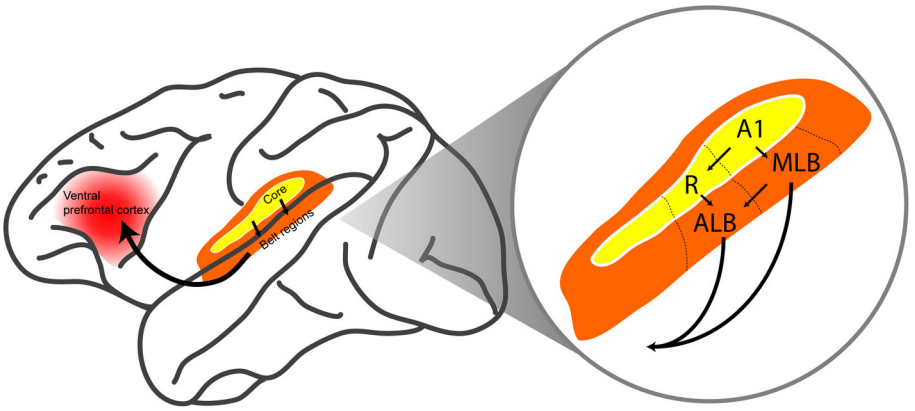


FIGURE 1 | The ventral auditory pathway in the monkey brain. The ventral auditory pathway begins in core auditory cortex (in particular, the primary auditory cortex A1 and the rostral field R). The pathway continues into the middle-lateral (MLB) and anterolateral (ALB) belt regions, which project directly and indirectly to the ventral prefrontal cortex. Arrows indicate feedforward projections. The figure is modified, with permission, from Hackett et al. (1998) and Romanski et al. (1999a).

These observations are consistent with the idea that there is a progression of category-information processing along the ventral auditory pathway: brain regions become increasingly sensitive to more complex types of categories. More specifically, it appears that neurons in core auditory cortex may encode categories for simple sounds, whereas neurons in the belt regions and the ventral prefrontal cortex may encode categories for more complex sounds and abstract information.

Indeed, neural correlates of auditory categorization can be seen in the core auditory cortex for simple frequency contours (Ohl et al., 2001; Selezneva et al., 2006). For example, in a study by Selezneva and colleagues, monkeys categorized the direction of a frequency contour of tone-burst sequences as either "increasing" or "decreasing" while neural activity was recorded from the primary auditory cortex. Selezneva et al. found that these core neurons encoded the sequence direction independent of its specific frequency content: that is, a core neuron responded similarly to a decreasing sequence from 1 to 0.5 kHz as it did to a decreasing sequence from 6 to 3 kHz. In a second study, Ohl et al. demonstrated that categorical representations need not be represented in the firing rates of single neurons but, instead, can be encoded in the dynamic firing patterns of a neural population. Thus, even in the earliest stage of the ventral auditory pathway, there is evidence for neural categorization.

Although the core auditory cortex processes categorical information for simple auditory stimuli (e.g., the direction of frequency changes of pure tones), studies using more complex sounds, such as human-speech sounds, have shown that core neurons primarily encode the acoustic features that compose these complex sounds but do not encode their category membership (Liebenthal et al., 2005; Steinschneider et al., 2005; Obleser et al., 2007; Engineer et al., 2008, 2013; Mesgarani et al., 2008, 2014; Nourski et al., 2009; Steinschneider, 2013). That is, the categorization of complex sounds requires not only analyses at the level of the acoustic feature but also subsequent computations that integrate the analyzed features into a perceptual representation, which is then subject to a categorization process. For example, distributed and temporally dynamic neural responses in individual core neurons can represent different acoustic features of speech sounds (Schreiner, 1998; Steinschneider et al., 2003; Engineer et al., 2008; Mesgarani et al., 2008, 2014), but the categorization of the speech sounds requires classifying the activation pattern across the entire population of core neurons.

Categorical representations of speech sounds at the level of the single neuron or local populations of neurons appear to occur at the next stage of auditory processing in the ventral auditory pathway, the lateral-belt regions. Several recent studies have noted that neural activity in the monkey lateral-belt and human superior temporal gyrus encodes speech-sound categories (Chang et al., 2010; Steinschneider et al., 2011; Tsunada et al., 2011; Steinschneider, 2013). For example, our group found that, when monkeys categorized two prototypes of speech sounds ("bad" and "dad") and their morphed versions, neural activity in the lateral belt discretely changed at the category boundary, suggesting that these neurons encoded the auditory category rather than smoothly varying acoustic features (**Figure 2**).

Human-neuroimaging studies have also found that the superior temporal sulcus is categorically activated by speech sounds, relative to other sounds (Binder et al., 2000; Leaver and Rauschecker, 2010). Specifically, the superior temporal sulcus was activated more by speech sounds than by frequency-modulated tones (Binder et al., 2000) or by other sounds including bird songs and animal vocalizations (Leaver and Rauschecker, 2010). Furthermore, activity in the superior temporal sulcus did not simply reflect the acoustic properties of speech sounds but, instead, represented the perception of speech (Mottonen et al., 2006; Desai et al., 2008).

Additionally, studies with other complex stimuli provide further evidence for the categorical encoding of complex sounds in the human non-primary auditory cortex, including superior temporal gyrus and sulcus, but not in the core auditory cortex (Altmann et al., 2007; Doehrmann et al., 2008; Leaver and Rauschecker, 2010). These studies found that complex sound categories were represented in spatially distinct and widely

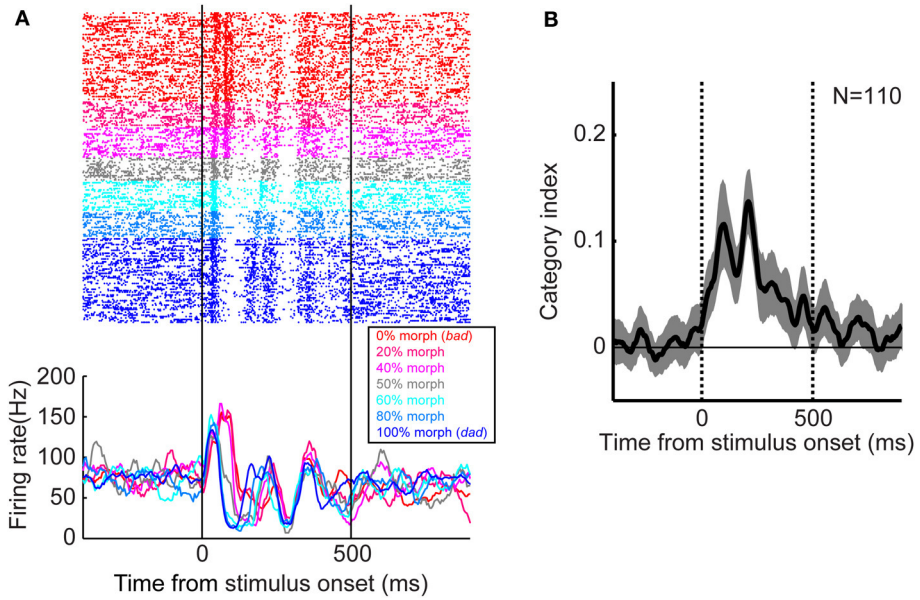


FIGURE 2 | Categorical neural activity in the monkey lateral belt during categorization of speech sounds. (A) An example of the activity of a lateral belt neuron. The speech sounds were two human-speech sounds ("bad" and "dad") and their morphs. Neural activity is color-coded by morphing percentage of the stimulus as shown in the legend. The raster plots and histograms are aligned relative to onset of the stimulus. **(B)** Temporal dynamics of the category index at the population level. Category-index values >0 indicate that neurons categorically represent speech sounds (Freedman et al., 2001; Tsunada et al., 2011). The thick line represents the mean value and the shaded area represents the bootstrapped 95%-confidence intervals of the mean. The two vertical lines indicate stimulus onset and offset, respectively, whereas the horizontal line indicates a category-index value of 0. The figure is adopted, with permission, from Tsunada et al. (2011).

distributed sub-regions within the superior temporal gyrus and sulcus (Obleser et al., 2006, 2010; Engel et al., 2009; Staeren et al., 2009; Chang et al., 2010; Leaver and Rauschecker, 2010; Giordano et al., 2013). For example, distinct regions of the superior temporal gyrus and sulcus are selectively activated by musical-instrument sounds (Leaver and Rauschecker, 2010), tool sounds (Doehrmann et al., 2008), and human-speech sounds (Belin et al., 2000; Binder et al., 2000; Warren et al., 2006); whereas the anterior part of the superior temporal gyrus and sulcus is preferentially activated by the passive listening of conspecific vocalizations than other vocalizations (Fecteau et al., 2004). Similar findings for con-specific vocalizations have been obtained in the monkey auditory cortex (Petkov et al., 2008; Perrodin et al., 2011). Consistent with these findings, neuropsychological studies have shown that human patients with damage in the temporal cortex have deficits in voice recognition and discrimination (i.e., phonagnosia Van Lancker and Canter, 1982; Van Lancker et al., 1988; Goll et al., 2010). Thus, hierarchically higher regions in the auditory cortex encode complex-sound categories in spatially distinct (i.e., modular) and widely distributed sub-regions.

Moreover, recent studies posit that the sub-regions in the non-primary auditory cortex process categorical information in a hierarchical manner (Warren et al., 2006). A recent meta-analysis of human speech-processing studies suggests that a hierarchical organization of speech processing exists within the superior temporal gyrus: the middle superior temporal gyrus is sensitive to phonemes; anterior superior temporal gyrus to words; and the most anterior locations to short phrases (Dewitt and Rauschecker, 2012; Rauschecker, 2012). Additionally, a different hierarchical processing of speech sounds in the superior temporal sulcus has also been articulated: the posterior superior temporal sulcus is preferentially sensitive for newly acquired sound categories, whereas the middle and anterior superior temporal sulci are more responsive to familiar sound categories (Liebenthal et al., 2005, 2010). Thus, within different areas of the non-primary auditory cortex, multiple and parallel processing may progress during auditory categorization.

Beyond the auditory cortex, do latter processing stages (e.g., the monkey ventral prefrontal cortex and human inferior frontal cortex) process categories for even more complex sounds? A re-examination of previous findings from our lab (Russ et al., 2008b; Tsunada et al., 2011) indicated important differences in neural categorization between the lateral belt and the ventral prefrontal cortex (**Figure 3**). We found that, at the population level, the category sensitivity for speech sounds in the prefrontal cortex was weaker than that in the lateral belt although neural activity in the prefrontal cortex transmitted a significant amount of categorical information. Consistent with this finding, a human-neuroimaging study also found that neural activity in the superior temporal gyrus is better correlated with a listener's ability to discriminate between speech sounds than the activity in the inferior prefrontal cortex (Binder et al., 2004). Because complex sounds, including speech sounds, are substantially processed in the non-primary auditory cortex as discussed above, the prefrontal cortex may not represent, relative to the auditory cortex, a higher level of auditory perceptual-feature categorization.

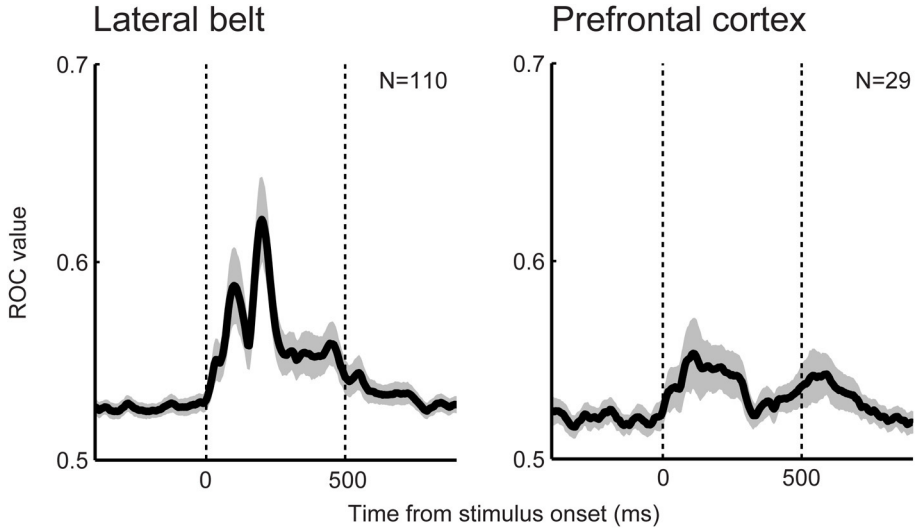


FIGURE 3 | Category sensitivity for speech sounds in the prefrontal cortex (right) is weaker than that in the lateral belt (left). Temporal dynamics of the category sensitivity at the population level are shown. Category sensitivity was calculated using a receiver-operating-characteristic (ROC) analysis (Green and Swets, 1966; Tsunada et al., 2012). Larger ROC values indicate better differentiation between the two categories.

Instead, the prefrontal cortex may be more sensitive to categories that are formed based on the abstract information that is transmitted by sounds. For example, the human inferior prefrontal cortex may encode categories for abstract information like emotional valence of a speaker's voice (Fecteau et al., 2005). Furthermore, human electroencephalography and neuroimaging studies have also revealed that the inferior prefrontal cortex plays a key role in the categorization of semantic information of multisensory stimuli (Werner and Noppeney, 2010; Joassin et al., 2011; Hu et al., 2012): Joassin et al. showed that the inferior prefrontal cortex contains multisensory category representations of gender that is derived from a speaker's voice and from visual images of a person's face.

Similarly, the monkey ventral prefrontal cortex encodes abstract categories. We have found that neurons in the ventral prefrontal cortex represent categories for food-related calls based on the transmitted information (e.g., high quality food vs. low quality food) (Gifford et al., 2005; Cohen et al., 2006). A more recent study found that neural activity in the monkey prefrontal cortex categorically represents the number of auditory stimuli (Nieder, 2012). Thus, along the ascending auditory system in the ventral auditory pathway, cortical areas encode categories for more complex stimuli and more abstract information.

NEURAL TRANSFORMATIONS WITHIN LOCAL MICROCIRCUITS

In this section, we discuss how the categorical information represented in each cortical area of the ventral auditory pathway is computed within local microcircuits. First, we briefly review the cortical microcircuit. Next, we focus on the role that two main cell classes of neurons in cortical microcircuits (i.e., excitatory pyramidal neurons and inhibitory interneurons) and discuss how different classes of neurons process categorical information.

HOW DO DIFFERENT CLASSES OF NEURONS IN LOCAL MICROCIRCUITS PROCESS CATEGORICAL INFORMATION?

A cortical microcircuit can be defined as a functional unit that processes inputs and generates outputs by dynamic and local interactions of excitatory pyramidal neurons and inhibitory interneurons (Merchant et al., 2012). Consequently, pyramidal neurons and interneurons are considered to be the main elements of microcircuits. Pyramidal neurons, which consist ∼70–90% of cortical neurons, provide excitatory-outputs locally (i.e., within a cortical area) and across brain areas (Markham et al., 2004). On the other hand, interneurons, which consist small portion of cortical neurons (∼10–30%), provide mainly inhibitory-outputs to surrounding pyramidal neurons and other interneurons (Markham et al., 2004).

From a physiological perspective, pyramidal neurons and interneurons can be classified based on the waveform of their action potentials (Mountcastle et al., 1969; McCormick et al., 1985; Kawaguchi and Kubota, 1993, 1997; Kawaguchi and Kondo, 2002; Markham et al., 2004; González-Burgos et al., 2005). More specifically, the waveforms of pyramidal neurons tend to be broader and slower than those seen in the most interneurons. Using this classification, several extracellular-recording studies have been able to elucidate roles of pyramidal neurons and interneurons for visual working memory in the prefrontal cortex (Wilson et al., 1994; Rao et al., 1999; Constantinidis and Goldman-Rakic, 2002; Diester and Nieder, 2008; Hussar and Pasternak, 2012), visual attention in V4 (Mitchell et al., 2007), visual perceptual decision-making in the frontal eye field (Ding and Gold, 2011), motor control in the motor and premotor cortices (Isomura et al., 2009; Kaufman et al., 2010), and auditory processing during the passive listening in the auditory cortex (Atencio and Schreiner, 2008; Sakata and Harris, 2009; Ogawa et al., 2011). Interestingly, most of these studies showed differential roles in pyramidal neurons and interneurons.

Recently, using differences in the waveform of extracellularly-recorded neurons, we found that putative pyramidal neurons and interneurons in the lateral belt differentially encode and represent auditory categories (Tsunada et al., 2012). Specifically, we found that interneurons, on average, are more sensitive for auditory-category information than pyramidal neurons, although both neuron classes reliably encode category information (**Figure 4**).

Unfortunately, to our knowledge, there have not been other auditory-category studies that have examined the relative category sensitivity of pyramidal neurons vs. interneurons. However, a comparable visual-categorization study on numerosity in the prefrontal cortex (Diester and Nieder, 2008) provides an opportunity to compare results across studies. Unlike our finding, Diester and Nieder found greater category sensitivity for putative pyramidal neurons than for putative interneurons.

The bases for these different sets of findings are unclear. However, three non-exclusive possibilities may underlie these differences. One possibility may relate to differences in the local-connectivity patterns and interactions between pyramidal neurons and interneurons across cortical areas (Wilson et al., 1994; Constantinidis and Goldman-Rakic, 2002; Diester and Nieder, 2008; Kätzel et al., 2010; Tsunada et al., 2012). Indeed, in the prefrontal cortex, simultaneously recorded (and, hence, nearby)

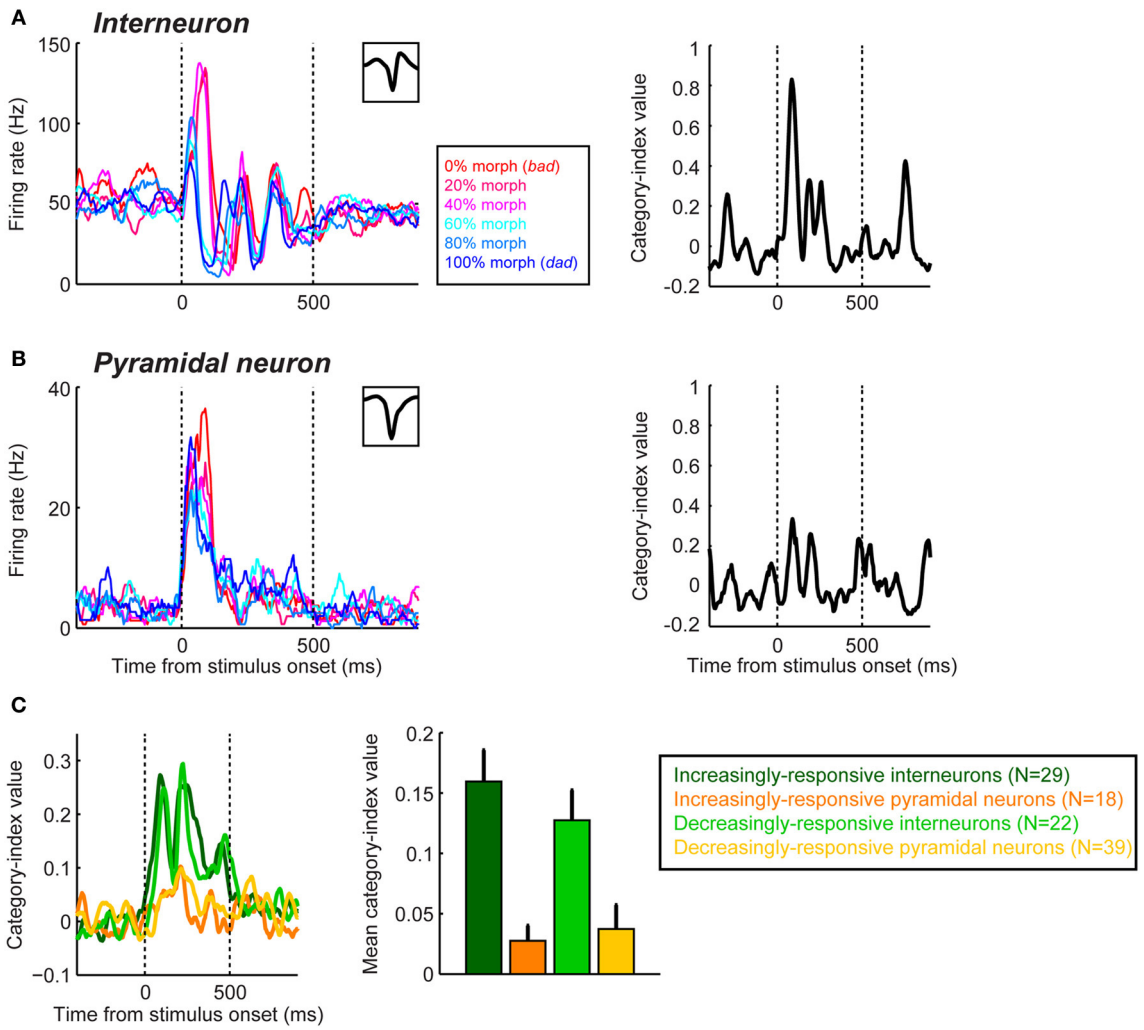


FIGURE 4 | Category sensitivity in interneurons is greater than that seen in pyramidal neurons during categorization of speech sounds in the auditory cortex. The plots in the left column of panel **(A,B)** show the mean firing rates of an interneuron **(A)** and a pyramidal neuron **(B)** as a function of time and the stimulus presented. The stimuli were two human-speech sounds ("bad" and "dad") and their morphs. Neural activity is color-coded by morphing percentage of the stimulus as shown in the legend. The inset in the upper graph of each plot shows the neuron's spike-waveform. The right column shows each neuron's category-index values as a function of time. For all of the panels, the two vertical dotted lines indicate stimulus onset and offset, respectively. **(C)** Population results of category index. The temporal profile (left panel) and mean (right) of the category index during the stimulus presentation are shown. Putative interneurons and pyramidal neurons were further classified as either "increasingly responsive" or "decreasingly responsive" based on their auditory-evoked responses. Error bars represent bootstrapped 95% confidence intervals of the mean. The figure is adopted, with permission, from Tsunada et al. (2012).

pyramidal neurons and interneurons have different category preferences (Diester and Nieder, 2008). In contrast, in the auditory cortex, simultaneously recorded pairs of pyramidal neurons and interneurons have similar category preferences (Tsunada et al., 2012). Thus, there may be different mechanisms for shaping category sensitivity across cortical areas. Second, the nature of the categorization task may also affect, in part, the category sensitivity of pyramidal neurons and interneurons: our task was a relatively simple task requiring the categorization of speech sounds based primarily on perceptual similarity, whereas Diester and Nieder's study required a more abstract categorization of numerosity. Finally, the third possibility relates to differences between stimulus dynamics: the visual stimuli in the Diester and Nieder's study were static stimuli, whereas our speech sounds had a rich spectrotemporal dynamic structure. To categorize dynamic stimuli, the moment-by-moment features of stimuli need to be quickly categorized. Thus, the greater category sensitivity of interneurons along with their well-known inhibitory influence on pyramidal neurons (Hefti and Smith, 2003; Wehr and Zador, 2003; Atencio and Schreiner, 2008; Fino and Yuste, 2011; Isaacson and Scanziani, 2011; Packer and Yuste, 2011; Zhang et al., 2011) may underlie the neural computations needed to create categorical representations of dynamic stimuli in the auditory cortex.

CONCLUSIONS AND FUTURE DIRECTIONS

Different neural transformations across different scales of neural organization progress during auditory categorization. Along the ascending auditory system in the ventral pathway, there is a progression in the encoding of categories from simple acoustic categories to categories representing abstract information. On the other hand, in local microcircuits within a cortical area, different classes of neurons, pyramidal neurons and interneurons, differentially compute categorical information. The computation is likely dependent upon the functional organization of the cortical area and dynamics of stimuli.

Despite several advances in our understanding of neural mechanism of auditory categorization, there still remain many important questions to be addressed. For example, it is poorly understood how bottom-up inputs from hierarchically lower areas, top-down feedback from higher areas, and local computations interact to form neural representations of auditory categories. Answering this question will provide a more thorough understanding of the information flow in the ventral auditory pathway. Another important question to be tested is what neural circuit mechanisms produce different category sensitivity between pyramidal neurons and interneurons, and functional roles of pyramidal neurons and interneurons in auditory categorization. Relevant to this question, the role that cortical laminae (another key element of local microcircuitry) play in auditory categorization should be also tested. Recent advances in experimental and analysis techniques should enable us to clarify the functional role of different classes of neurons in auditory categorization (Letzkus et al., 2011; Znamenskiy and Zador, 2013) and also test neural categorization across cortical layers (Lakatos et al., 2008; Takeuchi et al., 2011), providing further insights for neural computations for auditory categorization within local microcircuits.

ACKNOWLEDGMENTS

We thank Kate Christison-Lagay, Steven Eliades, and Heather Hersh for helpful comments on the preparation of this manuscript. We also thank Brian Russ and Jung Lee for data collection and Harry Shirley for outstanding veterinary support in our previous experiments. Joji Tsunada and Yale E. Cohen were supported by grants from NIDCD-NIH and the Boucai Hearing Restoration Fund.

REFERENCES

Altmann, C. F., Doehrmann, O., and Kaiser, J. (2007). Selectivity for animal vocalizations in the human auditory cortex. *Cereb. Cortex* 17, 2601–2608. doi: 10.1093/cercor/bhl167

Atencio, C. A., and Schreiner, C. E. (2008). Spectrotemporal processing differences between auditory cortical fast-spiking and regular-spiking neurons. *J. Neurosci.* 28, 3897–3910. doi: 10.1523/JNEUROSCI.5366-07.2008

Averbeck, B. B., and Romanski, L. M. (2004). Principal and independent components of macaque vocalizations: constructing stimuli to probe high-level sensory processing. *J. Neurophysiol.* 91, 2897–2909. doi: 10.1152/jn.01103.2003

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–311. doi: 10.1038/35002078

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

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

Bizley, J. K., and Cohen, Y. E. (2013). The what, where, and how of auditory-object perception. *Nat. Rev. Neurosci.* 14, 693–707. doi: 10.1038/nrn3565

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Cohen, Y. E., Hauser, M. D., and Russ, B. E. (2006). Spontaneous processing of abstract categorical information in the ventrolateral prefrontal cortex. *Biol. Lett.* 2, 261–265. doi: 10.1098/rsbl.2005.0436

Cohen, Y. E., Russ, B. E., Davis, S. J., Baker, A. E., Ackelson, A. L., and Nitecki, R. (2009). A functional role for the ventrolateral prefrontal cortex in non-spatial auditory cognition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20045–20050. doi: 10.1073/pnas.0907248106

Cohen, Y. E., Theunissen, F., Russ, B. E., and Gill, P. (2007). Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex. *J. Neurophysiol.* 97, 1470–1484. doi: 10.1152/jn.00769.2006

Constantinidis, C., and Goldman-Rakic, P. S. (2002). Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *J. Neurophysiol.* 88, 3487–3497. doi: 10.1152/jn.00188.2002

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

Dewitt, 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

Dicarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Diester, I., and Nieder, A. (2008). Complementary contributions of prefrontal neuron classes in abstract numerical categorization. *J. Neurosci.* 28, 7737–7747. doi: 10.1523/JNEUROSCI.1347-08.2008

Ding, L., and Gold, J. I. (2011). Neural correlates of perceptual decision making before, during, and after decision commitment in monkey frontal eye field. *Cereb. Cortex* 22, 1052–1067. doi: 10.1093/cercor/bhr178

Doehrmann, O., Naumer, M. J., Volz, S., Kaiser, J., and Altmann, C. F. (2008). Probing category selectivity for environmental sounds in the human auditory brain. *Neuropsychologia* 46, 2776–2786. doi: 10.1016/j.neuropsychologia.2008.05.011

Eckardt, W., and Zuberbuhler, K. (2004). Cooperation and competition in two forest monkeys. *Behav. Ecol.* 15, 400–411. doi: 10.1093/beheco/arh032

Engel, L. R., Frum, C., Puce, A., Walker, N. A., and Lewis, J. W. (2009). Different categories of living and non-living sound-sources activate distinct cortical networks. *Neuroimage* 47, 1778–1791. doi: 10.1016/j.neuroimage.2009.05.041

Engineer, C. T., Perez, C. A., Carraway, R. S., Chang, K. Q., Roland, J. L., Sloan, A. M., et al. (2013). Similarity of cortical activity patterns predicts generalization behavior. *PLoS ONE* 8:e78607. doi: 10.1371/journal.pone.0078607

Engineer, C. T., Perez, C. A., Chen, Y. H., Carraway, R. S., Reed, A. C., Shetake, J. A., et al. (2008). Cortical activity patterns predict speech discrimination ability. *Nat. Neurosci.* 11, 603–608. doi: 10.1038/nn.2109

Fecteau, S., Armony, J. L., Joanette, Y., and Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *Neuroimage* 23, 840–848. doi: 10.1016/j.neuroimage.2004.09.019

Fecteau, S., Armony, J. L., Joanette, Y., and Belin, P. (2005). Sensitivity to voice in human prefrontal cortex. *J. Neurophysiol.* 94, 2251–2254. doi: 10.1152/jn.00329.2005

Fino, E., and Yuste, R. (2011). Dense inhibitory connectivity in neocortex. *Neuron* 69, 1188–11203. doi: 10.1016/j.neuron.2011.02.025

Freedman, D. J., and Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature* 443, 85–88. doi: 10.1038/nature05078

Freedman, D. J., and Miller, E. K. (2008). Neural mechanisms of visual categorization: insights from neurophysiology. *Neurosci. Biobehav. Rev.* 32, 311–329. doi: 10.1016/j.neubiorev.2007.07.011

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316. doi: 10.1126/science.291.5502.312

Gifford, G. W. 3rd., Maclean, K. A., Hauser, M. D., and Cohen, Y. E. (2005). The neurophysiology of functionally meaningful categories: macaque ventrolateral prefrontal cortex plays a critical role in spontaneous categorization of species-specific vocalizations. *J. Cogn. Neurosci.* 17, 1471–1482. doi: 10.1162/0898929054985464

Gifford, G. W. 3rd., Hauser, M. D., and Cohen, Y. E. (2003). Discrimination of functionally referential calls by laboratory-housed rhesus macaques: implications for neuroethological studies. *Brain Behav. Evol.* 61, 213–224. doi: 10.1159/000070704

Giordano, B. L., McAdams, S., Kriegeskorte, N., Zatorre, R. J., and Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cereb. Cortex* 23, 2025–2037. doi: 10.1093/cercor/bhs162

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron* 14, 477–485. doi: 10.1016/0896-6273(95)90304-6

Goll, J. C., Crutch, S. J., and Warren, J. D. (2010). Central auditory disorders: toward a neuropsychology of auditory objects. *Curr. Opin. Neurol.* 23, 617–627. doi: 10.1097/WCO.0b013e32834027f6

González-Burgos, G., Krimer, L. S., Povysheva, N. V., Barrionuevo, G., and Lewis, D. A. (2005). Functional properties of fast spiking interneurons and their synaptic connections with pyramidal cells in primate dorsolateral prefrontal cortex. *J. Neurophysiol.* 93, 942–953. doi: 10.1152/jn.00787.2004

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York, NY: John Wiley and Sons, Inc.

Hackett, T. A., Stepniewska, I., and Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *J. Comp. Neurol.* 394, 475–495. doi: 10.1002/(SICI)1096-9861(19980518)394:4<475::AID-CNE6>3.0.CO;2-Z

Hauser, M. D. (1998). Functional referents and acoustic similarity: field playback experiments with rhesus monkeys. *Anim. Behav.* 55, 1647–1658. doi: 10.1006/anbe.1997.0712

Hauser, M. D., and Marler, P. (1993a). Food-associated calls in rhesus macaques (*Macaca mulatta*) 1. Socioecological factors influencing call production. *Behav. Ecol.* 4, 194–205. doi: 10.1093/beheco/4.3.194

Hauser, M. D., and Marler, P. (1993b). Food-associated calls in rhesus macaques (*Macaca mulatta*) II. Costs and benefits of call production and suppression. *Behav. Ecol.* 4, 206–212. doi: 10.1093/beheco/4.3.206

Hefti, B. J., and Smith, P. H. (2003). Distribution and kinetic properties of GABAergic inputs to layer V pyramidal cells in rat auditory cortex. *J. Assoc. Res. Otolaryngol.* 4, 106–121. doi: 10.1007/s10162-002-3012-z

Holt, L. L., and Lotto, A. J. (2010). Speech perception as categorization. *Atten. Percept. Psychophys.* 72, 1218–1227. doi: 10.3758/APP.72.5.1218

Hussar, C. R., and Pasternak, T. (2012). Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. *J. Neurosci.* 32, 2747–2761. doi: 10.1523/JNEUROSCI.5135-11.2012

Hu, Z., Zhang, R., Zhang, Q., Liu, Q., and Li, H. (2012). Neural correlates of audiovisual integration of semantic category information. *Brain Lang.* 121, 70–75. doi: 10.1016/j.bandl.2012.01.002

Isaacson, J. S., and Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron* 72, 231–243. doi: 10.1016/j.neuron.2011.09.027

Isomura, Y., Harukuni, R., Takekawa, T., Aizawa, H., and Fukai, T. (2009). Microcircuitry coordination of cortical motor information in self-initiation of voluntary movements. *Nat. Neurosci.* 12, 1586–1593. doi: 10.1038/nn.2431

Joassin, F., Maurage, P., and Campanella, S. (2011). The neural network sustaining the crossmodal processing of human gender from faces and voices: an fMRI study. *Neuroimage* 54, 1654–1661. doi: 10.1016/j.neuroimage.2010.08.073

Kaas, J. H., and Hackett, T. A. (1999). "What" and "where" processing in auditory cortex. *Nat. Neurosci.* 2, 1045–1047. doi: 10.1038/15967

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799. doi: 10.1073/pnas.97.22.11793

Kaas, J. H., Hackett, T. A., and Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Curr. Opin. Neurobiol.* 9, 164–170. doi: 10.1016/S0959-4388(99)80022-1

Kätzel, D., Zemelman, B. V., Buetfering, C., Wölfel, M., and Miesenböck, G. (2010). The columnar and laminar organization of inhibitory connections to neocortical excitatory cells. *Nat. Neurosci.* 14, 100–107. doi: 10.1038/nn.2687

Kaufman, M. T., Churchland, M. M., Santhanam, G., Yu, B. M., Afshar, A., Ryu, S. I., et al. (2010). Roles of monkey premotor neuron classes in movement preparation and execution. *J. Neurophysiol.* 104, 799–810. doi: 10.1152/jn.00231.2009

Kawaguchi, Y., and Kondo, S. (2002). Parvalbumin, somatostatin and cholecystokinin as chemical markers for specific GABAergic interneuron types in the rat frontal cortex. *J. Neurocytol.* 31, 277–287. doi: 10.1023/A:1024126110356

Kawaguchi, Y., and Kubota, Y. (1993). Correlation of physiological subgroupings of nonpyramidal cells with parvalbumin- and calbindinD28k-immunoreactive neurons in layer V of rat frontal cortex. *J. Neurophysiol.* 70, 387–396.

Kawaguchi, Y., and Kubota, Y. (1997). GABAergic cell subtypes and their synaptic connections in rat frontal cortex. *Cereb. Cortex* 7, 476–486. doi: 10.1093/cercor/7.6.476

Kluender, K. R., Diehl, R. L., and Killeen, P. (1987). Japanese quail can learn phonetic categories. *Science* 237, 1195–1197. doi: 10.1126/science.3629235

Kuhl, P. K., and Miller, J. D. (1975). Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. *Science* 190, 69–72. doi: 10.1126/science.1166301

Kuhl, P. K., and Miller, J. D. (1978). Speech perception by the chinchilla: identification function for synthetic VOT stimuli. *J. Acoust. Soc. Am.* 63, 905–917. doi: 10.1121/1.381770

Kuhl, P. K., and Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Percept. Psychophys.* 32, 542–550. doi: 10.3758/BF03204208

Kuhl, P. K., and Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J. Acoust. Soc. Am.* 73, 1003–1010. doi: 10.1121/1.389148

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

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Letzkus, J. J., Wolff, S. B., Meyer, E. M., Tovote, P., Courtin, J., Herry, C., et al. (2011). A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* 480, 331–335. doi: 10.1038/nature10674

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 5, 552–563. doi: 10.1037/h0020279

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Lotto, A. J., Kluender, K. R., and Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *J. Acoust. Soc. Am.* 102, 1134–1140. doi: 10.1121/1.419865

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Percept. Psychophys.* 28, 407–412. doi: 10.3758/BF03204884

Markham, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneuron of the neocortical inhibitory system. *Nat. Rev. Neurosci.* 5, 793–807. doi: 10.1038/nrn1519

McCormick, D. A., Connors, B. W., Lighthall, J. W., and Prince, D. A. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol.* 54, 782–806.

Merchant, H., De Lafuente, V., Pena-Ortega, F., and Larriva-Sahd, J. (2012). Functional impact of interneuronal inhibition in the cerebral cortex of behaving animals. *Prog. Neurobiol.* 99, 163–178. doi: 10.1016/j.pneurobio.2012.08.005

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.* 123, 899–909. doi: 10.1121/1.2816572

Miller, C. T., and Cohen, Y. E. (2010). "Vocalization processing," in *Primate Neuroethology*, eds A. Ghazanfar and M. L. Platt (Oxford, UK: Oxford University Press), 237–255. doi: 10.1093/acprof:oso/9780195326598.003.0013

Miller, E., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* 1, 59–65. doi: 10.1038/35036228

Miller, E. K., Freedman, D. J., and Wallis, J. D. (2002). The prefrontal cortex: categories, concepts, and cognition. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 29, 1123–1136. doi: 10.1098/rstb.2002.1099

Miller, E. K., Nieder, A., Freedman, D. J., and Wallis, J. D. (2003). Neural correlates of categories and concepts. *Curr. Opin. Neurobiol.* 13, 198–203. doi: 10.1016/S0959-4388(03)00037-0

Mitchell, J. F., Sundberg, K. A., and Reynolds, J. H. (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* 5, 131–141. doi: 10.1016/j.neuron.2007.06.018

Mottonen, R., Calvert, G. A., Jaaskelainen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage* 30, 563–569. doi: 10.1016/j.neuroimage.2005.10.002

Mountcastle, V. B., Talbot, W. H., Sakata, H., and Hyvarinen, J. (1969). Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys. Neuronal periodicity and frequency discrimination. *J. Neurophysiol.* 32, 452–484.

Nieder, A. (2012). Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11860–11865. doi: 10.1073/pnas.1204580109

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

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., et al. (2006). Vowel sound extraction in anterior superior temporal cortex. *Hum. Brain Mapp.* 27, 562–571. doi: 10.1002/hbm.20201

Obleser, J., Leaver, A. M., Van Meter, J., and Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Front. Psychol.* 1:232. doi: 10.3389/fpsyg.2010.00232

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cereb. Cortex* 17, 2251–2257. doi: 10.1093/cercor/bhl133

Ogawa, T., Riera, J., Goto, T., Sumiyoshi, A., Noaka, H., Jerbi, K., et al. (2011). Large-scale heterogeneous representation of sound attributes in rat primary auditory cortex: from unit activity to population dynamics. *J. Neurosci.* 31, 14639–14653. doi: 10.1523/JNEUROSCI.0086-11.2011

Ohl, F. W., Scheich, H., and Freeman, W. J. (2001). Change in pattern of ongoing cortical activity with auditory category learning. *Nature* 412, 733–736. doi: 10.1038/35089076

Osada, T., Adachi, Y., Kimura, H. M., and Miyashita, Y. (2008). Towards understanding of the cortical network underlying associative memory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 2187–2199. doi: 10.1098/rstb.2008.2271

Packer, A. M., and Yuste, R. (2011). Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? *J. Neurosci.* 14, 13260–13271. doi: 10.1523/JNEUROSCI.3131-11.2011

Pastore, R. E., Li, X. F., and Layer, J. K. (1990). Categorical perception of nonspeech chirps and bleats. *Percept. Psychophys.* 48, 151–156. doi: 10.3758/BF03207082

Perrodin, C., Kayser, C., Logothetis, N. K., and Petkov, C. I. (2011). Voice cells in the primate temporal lobe. *Curr. Biol.* 21, 1408–1415. doi: 10.1016/j.cub.2011.07.028

Petkov, C. I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., and Logothetis, N. K. (2008). A voice region in the monkey brain. *Nat. Neurosci.* 11, 367–374. doi: 10.1038/nn2043

Plakke, B., Diltz, M. D., and Romanski, L. M. (2013a). Coding of vocalizations by single neurons in ventrolateral prefrontal cortex. *Hear. Res.* 305, 135–143. doi: 10.1016/j.heares.2013.07.011

Plakke, B., Hwang, J., Diltz, M. D., and Romanski, L. M. (2013b). "The role of ventral prefrontal cortex in auditory, visual and audiovisual working memory," in *Society for Neuroscience*, *Program No. 574.515 Neuroscience Meeting Planner* (San Diego, CA).

Plakke, B., Ng, C. W., and Poremba, A. (2013c). Neural correlates of auditory recognition memory in primate lateral prefrontal cortex. *Neuroscience* 244, 62–76. doi: 10.1016/j.neuroscience.2013.04.002

Poremba, A., Bigelow, J., and Rossi, B. (2013). Processing of communication sounds: contributions of learning, memory, and experience. *Hear. Res.* 305, 31–44. doi: 10.1016/j.heares.2013.06.005

Rao, S. G., Williams, G. V., and Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: evidence for microcolumnar organization in PFC. *J. Neurophysiol.* 81, 1903–1915.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Curr. Opin. Neurobiol.* 8, 516–521. doi: 10.1016/S0959-4388(98)80040-8

Rauschecker, J. P. (2012). Ventral and dorsal streams in the evolution of speech and language. *Front. Evol. Neurosci.* 4:7. doi: 10.3389/fnevo.2012.00007

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

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111–114. doi: 10.1126/science.7701330

Recanzone, G. H., and Cohen, Y. E. (2010). Serial and parallel processing in the primate auditory cortex revisited. *Behav. Brain Res.* 5, 1–6. doi: 10.1016/j.bbr.2009.08.015

Romanski, L. M., and Averbeck, B. B. (2009). The primate cortical auditory system and neural representation of conspecific vocalizations. *Annu. Rev. Neurosci.* 32, 315–346. doi: 10.1146/annurev.neuro.051508.135431

Romanski, L. M., Bates, J. F., and Goldman-Rakic, P. S. (1999a). Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *J. Comp. Neurol.* 403, 141–157. doi: 10.1002/(SICI)1096-9861(19990111)403:2<141::AID-CNE1>3.0.CO;2-V

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999b). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136. doi: 10.1038/16056

Russ, B. E., Ackelson, A. L., Baker, A. E., and Cohen, Y. E. (2008a). Coding of auditory-stimulus identity in the auditory non-spatial processing stream. *J. Neurophysiol.* 99, 87–95. doi: 10.1152/jn.01069.2007

Russ, B. E., Lee, Y.-S., and Cohen, Y. E. (2007). Neural and behavioral correlates of auditory categorization. *Hear. Res.* 229, 204–212. doi: 10.1016/j.heares.2006.10.010

Russ, B. E., Orr, L. E., and Cohen, Y. E. (2008b). Prefrontal neurons predict choices during an auditory same-different task. *Curr. Biol.* 18, 1483–1488. doi: 10.1016/j.cub.2008.08.054

Sakata, S., and Harris, K. D. (2009). Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. *Neuron* 64, 404–418. doi: 10.1016/j.neuron.2009.09.020

Schreiner, C. E. (1998). Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. *Audiol. Neurootol.* 3, 104–122. doi: 10.1159/000013785

Selezneva, E., Scheich, H., and Brosch, M. (2006). Dual time scales for categorical decision making in auditory cortex. *Curr. Biol.* 16, 2428–2433. doi: 10.1016/j.cub.2006.10.027

Sinnott, J. M., and Brown, C. H. (1997). Perception of the American English liquid vertical bar ra-la vertical bar contrast by humans and monkeys. *J. Acoust. Soc. Am.* 102, 588–602. doi: 10.1121/1.419732

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.* 19, 498–502. doi: 10.1016/j.cub.2009.01.066

Steinschneider, M. (2013). "Phonemic representations and categories," in *Neural Correlates of Auditory Cognition*, eds Y. E. Cohne, A. N. Popper and R. R. Fay (New York, NY: Springer), 151–191. doi: 10.1007/978-1-4614-2350-8_6

Steinschneider, M., Fishman, Y. I., and Arezzo, J. C. (2003). Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. *J. Acoust. Soc. Am.* 114, 307–321. doi: 10.1121/1.1582449

Steinschneider, M., Nourski, K. V., Kawasaki, H., Oya, H., Brugge, J. F., and Howard, M. A. (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb. Cortex* 10, 2332–2347. doi: 10.1093/cercor/bhr014

Steinschneider, M., Volkov, I. O., Fishman, Y. I., Oya, H., Arezzo, J. C., and Howard, M. A. 3rd. (2005). Intracortical responses in human and monkey primary auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. *Cereb. Cortex* 15, 170–186. doi: 10.1093/cercor/bhh120

Takeuchi, D., Hirabayashi, T., Tamura, K., and Miyashita, Y. (2011). Reversal of interlaminar signal between sensory and memory processing in monkey temporal cortex. *Science* 331, 1443–1447. doi: 10.1126/science.1199967

Tian, B., and Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* 92, 2993–3013. doi: 10.1152/jn.00472.2003

Tsunada, J., Lee, J. H., and Cohen, Y. E. (2011). Representation of speech categories in the primate auditory cortex. *J. Neurophysiol.* 105, 2634–2646. doi: 10.1152/jn.00037.2011

Tsunada, J., Lee, J. H., and Cohen, Y. E. (2012). Differential representation of auditory categories between cell classes in primate auditory cortex. *J. Physiol.* 590, 3129–3139. doi: 10.1113/jphysiol.2012.232892

Van Lancker, D. R., and Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain Cogn.* 1, 185–195.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., and Dobkin, B. H. (1988). Phonagnosia—a dissociation between familiar and unfamiliar voices. *Cortex* 24, 195–209. doi: 10.1016/S0010-9452(88)80029-7

Warren, J. D., Scott, S. K., Price, C. J., and Griffiths, T. D. (2006). Human brain mechanisms for the early analysis of voices. *Neuroimage* 31, 1389–1397. doi: 10.1016/j.neuroimage.2006.01.034

Wehr, M. S., and Zador, A. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* 27, 442–446. doi: 10.1038/nature02116

Werner, S., and Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J. Neurosci.* 30, 2662–2675. doi: 10.1523/JNEUROSCI.5091-09.2010

Wilson, F. A., O'Scalaidhe, S. P., and Goldman-Rakic, P. S. (1994). Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 91, 4009–4013. doi: 10.1073/pnas.91.9.4009

Zhang, L. I., Zhou, Y., and Tao, H. W. (2011). Perspectives on: information and coding in mammalian sensory physiology: inhibitory synaptic mechanisms underlying functional diversity in auditory cortex. *J. Gen. Physiol.* 138, 311–320. doi: 10.1085/jgp.201110650

Znamenskiy, P., and Zador, A. M. (2013). Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* 497, 482–485. doi: 10.1038/nature12077

Züberbuhler, K. (2000a). Causal cognition in a non-human primate: field playback experiments with Diana monkeys. *Cognition* 76, 195–207. doi: 10.1016/S0010-0277(00)00079-2

Züberbuhler, K. (2000b). Interspecies semantic communication in two forest primates. *Proc. R. Soc. Lond. B Biol. Sci.* 267, 713–718. doi: 10.1098/rspb.2000.1061

Zuberbuhler, K., and Seyfarth, R. M. (1997). Diana monkey long-distance calls: messages for conspecifics and predators. *Anim. Behav.* 53, 589–604. doi: 10.1006/anbe.1996.0334

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: 14 March 2014; accepted: 27 May 2014; published online: 17 June 2014.
*Citation: Tsunada J and Cohen YE (2014) Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. Front. Neurosci. **8**:161. doi: 10.3389/fnins.2014.00161*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 17 November 2014
doi: 10.3389/fnins.2014.00374

Electrophysiological evidence for change detection in speech sound patterns by anesthetized rats

Piia Astikainen[1]*, Tanel Mällo[1], Timo Ruusuvirta[2,3] and Risto Näätänen[4,5,6]

[1] *Department of Psychology, University of Jyväskylä, Jyväskylä, Finland*
[2] *Centre for Learning Research, University of Turku, Turku, Finland*
[3] *Department of Teacher education/Rauma Unit, University of Turku, Rauma, Finland*
[4] *Institute of Psychology, University of Tartu, Tartu, Estonia*
[5] *Center of Functionally Integrative Neuroscience, University of Århus, Århus, Denmark*
[6] *Cognitive Brain Research Unit, Institute of Behavioral Sciences, University of Helsinki, Helsinki, Finland*

Edited by:
Lynne E. Bernstein, George Washington University, USA

Reviewed by:
Alexandra Bendixen, Carl von Ossietzky University of Oldenburg, Germany
Guangying Wu, George Washington University, USA

****Correspondence:***
Piia Astikainen, Department of Psychology, University of Jyväskylä, PO Box 35, Ylistönmäentie 33, 40014 Jyväskylä, Finland
e-mail: piia.astikainen@jyu.fi

Human infants are able to detect changes in grammatical rules in a speech sound stream. Here, we tested whether rats have a comparable ability by using an electrophysiological measure that has been shown to reflect higher order auditory cognition even before it becomes manifested in behavioral level. Urethane-anesthetized rats were presented with a stream of sequences consisting of three pseudowords carried out at a fast pace. Frequently presented "standard" sequences had 16 variants which all had the same structure. They were occasionally replaced by acoustically novel "deviant" sequences of two different types: structurally consistent and inconsistent sequences. Two stimulus conditions were presented for separate animal groups. In one stimulus condition, the standard and the pattern-obeying deviant sequences had an AAB structure, while the pattern-violating deviant sequences had an ABB structure. In the other stimulus condition, these assignments were reversed. During the stimulus presentation, local-field potentials were recorded from the dura, above the auditory cortex. Two temporally separate differential brain responses to the deviant sequences reflected the detection of the deviant speech sound sequences. The first response was elicited by both types of deviant sequences and reflected most probably their acoustical novelty. The second response was elicited specifically by the structurally inconsistent deviant sequences (pattern-violating deviant sequences), suggesting that rats were able to detect changes in the pattern of three-syllabic speech sound sequence (i.e., location of the reduplication of an element in the sequence). Since all the deviant sound sequences were constructed of novel items, our findings indicate that, similarly to the human brain, the rat brain has the ability to automatically generalize extracted structural information to new items.

Keywords: local-field potentials, pattern perception, auditory cortex, rat, mismatch negativity, speech

INTRODUCTION

The ability to detect abstract grammatical rules, i.e., principles that govern speech sound streams, is essential for learning a language. To investigate the infants' ability to extract abstract algebraic rules, Marcus et al. (1999) familiarized infants to sequences of syllables (or sentences) that followed a particular "grammatical" rule (e.g., "ga ti ga" for ABA). During the test, infants were observed to be more attentive to sequences that were grammatically inconsistent (e.g., "wo fe fe," which is ABB) than to those sequences that were consistent with grammatical rules (e.g., "wo fe wo"). Because the test sentences were different to those used in the training phase, the authors concluded that infants can extract an abstract rule and generalize it to novel instances. Also, detection of ABB and AAB structures were compared, and it was found that even if both structures have a reduplication element, the infants paid more attention to the inconsistent patterns.

It is not known, however, whether the ability to extract grammatical rules from speech sounds only applies to human linguistic cognition or whether this cognitive element has originally evolved for other, more general purposes. In the latter case, these skills could also be found in non-human animal species.

It is known that non-human animal species can process speech up to a certain level of cognitive complexity. Speech sound discrimination has been demonstrated in various animal species neurophysiologically (e.g., Dooling and Brown, 1990 in birds; Kraus et al., 1994 in guinea pigs, Ahmed et al., 2011 in rats), and on a behavioral level (e.g., Engineer et al., 2008 in rats; Sinnott et al., 1976 in monkeys; Sinnott and Mosteller, 2001 in gerbils). Also, word segmentation based on transitional probabilities has been demonstrated, on a behavioral level, in rats (Toro and Trobalon, 2005) as well as in cotton-top tamarins (Hauser et al., 2001). Extraction of grammatical rules (i.e., structural patterns) from speech sounds in non-human species has been studied in tamarin-monkeys and rats with similar stimulus conditions as applied originally by Marcus et al. (1999). The report concerning tamarin-monkeys (Hauser et al., 2002) was

later retracted (Retraction notice, 2010). In rats, no evidence of pattern extraction was found (Toro and Trobalon, 2005).

It might be, however, too early to conclude that rats are not able to extract structural patterns from three-syllabic speech sequences, as were applied in a classic study by Marcus et al. (1999) in infants. Since there is evidence in rats of representing abstract rules from pure tones (Murphy et al., 2008), this issue should be further explored. To the present study we applied a neurophysiological mismatch response (MMR), a measure of automatic cognition, which is the equivalent of the human electrophysiological response called mismatch negativity (MMN; Näätänen et al., 1978, 1997, 2010). MMR can reflect auditory cognition before its behavioral manifestation (e.g., Tremblay et al., 1998). Based on this method, we have previously demonstrated that the rat's brain is able to detect changes in abstract auditory features, such as melodic patterns in tone-pairs (Ruusuvirta et al., 2007) and in combinatory rules between frequency and intensity of the sound objects (Astikainen et al., 2006, 2014). Rats also make representations of spectro-temporally complex sounds such as speech sounds in their brains, and they can detect changes in these sounds based on the content of the transient memory (Ahmed et al., 2011). Rats, anesthetized with urethane have been used in these studies as urethane is known to largely preserve the awake-like function of the brain (Maggi and Meli, 1986).

In the present study, capitalizing on the above mentioned studies, we recorded local-field potentials (LFPs) from the dura, above the auditory cortex in urethane-anesthetized rats. We presented the animals with a series of synthesized speech sounds. The stimulus series (modified from Marcus et al., 1999) consisted of several different sequences consisting of three pseudowords (called sentences here). Ninety percent of the sentences followed a specific pattern structure ("standards"). Acoustically novel sentences were introduced ("deviants") rarely (10% of the sentences) and randomly in the sequences. Deviant sentences were of two different types: 1) "pattern-obeying deviants" that shared the pattern structure of the standard sentences but deviated from them physically, and 2) "pattern-violating deviants" that differed from the standards physically but also presented a different pattern structure. We expected to observe an early MMR to be triggered by the first pseudoword for both types of deviant sentences due to their acoustical differences from the standard pseudowords. We also expected to observe a later MMR to be triggered by the second word in the pattern-violating deviant sentences. This would indicate that the syntax-like rule, carried by the standard patterns, was extracted by the animals' brains.

MATERIALS AND METHODS

SUBJECTS

The subjects were 14 male Sprague-Dawley rats from Harlan Laboratories (England, UK), weighing 410–500 g and aged between 13 and 18 weeks at the time of the individual recordings. The animals were housed in standard plastic cages, in groups of 2–4, under a controlled temperature and subjected to a 12 h light/dark cycle, with free access to water and food pellets in the Experimental Animal Unit of the University of Jyväskylä, Jyväskylä, Finland. The experiments were approved by the Finnish National Animal Experiment Board, and carried out in accordance with the European Communities Council Directive (86/609/EEC) regarding the care and use of animals used for experimental procedures. The license for the present experiments has been approved by County Administrative Board of Southern Finland (Permit code: ESLH-2007-00662).

SURGERY

All surgical procedures were done under urethane (Sigma Chemicals, St Louis, MO, USA) induced anesthesia (1.2 g/kg dose, 0.24 g/ml concentration, injected intraperitoneally). Supplemental doses were injected if the required level of anesthesia was not obtained. The level of anesthesia was monitored by testing the withdrawal reflexes. The anesthetized animal was moved into a Faraday cage and mounted in a standard stereotactic frame (David Kopf Instruments, Model 962, Tujunga, CA, USA). The animal's head was fixed to the stereotaxic frame using blunt ear bars. Under additional local anesthesia (lidocaine 20%, Orion Pharma, Espoo, Finland), the skin was removed from the top of the head and the skull revealed. Positioned contralaterally to the recording site, two stainless steel skull screws (0.9 mm diameter, World Precision Instruments, Berlin, Germany) fixed above the cerebellum (AP −11.0, ML 3.0) and frontal cortex (AP +4.0, ML 3.0) served as reference and ground electrodes, respectively. A headstage, composed of a screw and dental acrylic, was attached to the right prefrontal part of the skull to hold the head in place and allow removal of the right ear bar. A unilateral craniotomy was performed in order to expose a 2 × 2 mm region over the left auditory cortex (4.5–6.5 mm posterior to the bregma and 2–4 mm lateral to the bony ridge between the dorsal and lateral skull surfaces) for the placement of the recording electrode. The level of anesthesia was periodically monitored throughout the whole experiment. Animals were rehydrated with a 2 ml injection of saline under the skin every 2 h. After the surgery, the right ear bar was removed and recording started. After the experiment, the animals were further anesthetized with urethane and then put down by cervical dislocation.

RECORDING

Local-field potentials in response to auditory stimuli were recorded with a teflon-coated stainless steel wire (200 μm in diameter, A-M Systems, Chantilly, VA) positioned on the dura surface above the left auditory cortex. Continuous electrocorticogram was primarily amplified 10-fold, by using the AI 405 amplifier (Molecular Devices Corporation, Union City, CA, USA), high-pass filtered at 0.1 Hz, 200-fold amplified, and low-pass filtered at 400 Hz (CyberAmp 380, Molecular Devices Corporation), and finally sampled with 16-bit precision at 2 kHz (DigiData 1320A, Molecular Devices Corporation). The data were stored on a computer hard disk using Axoscope 9.0 data acquisition software (Molecular Devices Corporation) for later off-line analysis.

STIMULI

Synthesized human male voice speech sounds which consisted of five formants, were created using Mikropuhe 5-software (Timehouse, Helsinki, Finland). The speech sound stream consisted of consonant-vowel syllables (words) that were 100 ms in

duration. These were presented in groups of three (modified from Marcus et al., 1999). There was a 50-ms pause between each consecutive word, within the sentences, and 100-ms pause between the sentences.

One of the two stimulus blocks (1 or 2) was presented in each animal ($n = 7$ for both blocks, see **Table 1**). In each block, 90% of the sentences ("standards") followed a specific structure. For one block, this structure was of AAB type (two identical words followed by a different word) and for the other block of ABB type (one word followed by two identical words). In each block, one structure was assigned to the standards (16 different variants, $p = 0.9$) and the other structure for the deviants ($p = 0.1$).The deviants were of two different types: (1) "pattern-obeying deviants" (2 variants, $p = 0.05$) that physically differed from the standards but obeyed the structure of standard sentences and (2) "pattern-violating deviants" (2 variants, $p = 0.05$) that differed from the standard sentences, both physically and in respect of the pattern. Since all the stimulus types included a repetition of an element, they were not possible to differentiate by detecting only this property of the stimulus. The sentences were ordered in a pseudorandom fashion with the restriction that consecutive deviants were separated by at least two standards. There were a total of 996 stimulus sequences in one stimulus block.

The speech sounds were played from a PC via an active loudspeaker system (Studiopro 3, M-audio, Irwindale, CA, USA). The stimulation was presented with the loudspeaker system directed toward the right ear of the animal at a distance of 20 cm. In all conditions, the sound pressure level for each tone was 70 dB, as measured with a sound level meter (type 2235, Bruel and Kjaer, Nærum Denmark) with C-weighting (optimized for 40–100 dB measurement) in the vicinity of the animal's right pinna during the recording.

ANALYSIS

The data were off-line filtered at 0.1–30 Hz (24 dB/octave roll off). Data of the two animal groups (stimulus blocks 1 and 2) were averaged. Sweeps from 50 ms before to 500 ms after each stimulus onset were segmented. In order to have same amount of standard and deviant responses in the analysis, only the responses to the standard sentences immediately preceding the deviant sentences were analyzed. The averaged waveforms were then baseline-corrected. The baseline correction was calculated for the period of -50 to 0 ms relative to the second word in the sentence since the change in the pattern occurred at that time in the pattern-violating deviants.

First, the timing of the MMR was investigated by applying point-by-point 2-tailed paired *t*-tests to compare local-field potential amplitudes for the standard and deviant sentences. *P*-values smaller than or equal to 0.05 for at least 20 consecutive sample points (i.e., for the period of 10 ms) were required for the difference in local-field potentials to be considered robust. Next, ANOVA with factors stimulus type (standard vs. deviant) and deviant type (pattern-obeying deviant vs. pattern-violating deviant) for the MMR specific to the pattern-violating deviant sentences was applied. For the ANOVA, mean amplitude values were extracted from the latency range of the significant differential response indicated by the point-by-point *t*-tests. Partial eta squared values present effect size estimates for ANOVA and Cohen's *d* for *t*-tests.

RESULTS

The first MMR, i.e., an amplitude difference in local-field potentials, between the standard and the deviant sentences, was found for both the pattern-violating deviant sentences (**Figure 1**, left) and the pattern-obeying deviant sentences (**Figure 1**, right). This first MMR for the pattern-violating deviant sentences, was significant at 194–213 ms after the sentence onset, [$t_{(13)} = 2.2$–2.7, $p = 0.020$–0.047], and at 231.5–251 ms after the sentence onset, [$t_{(13)} = 2.2$–2.3, $p = 0.039$–0.050]. For the pattern-obeying deviant sentences the corresponding latency ranges were 187.5–206.5 ms after the sentence onset, [$t_{(13)} = 2.155$–2.379, $p = 0.033$–0.050],

Table 1 | Stimulus categories and sequence variants.

	Stimulus categories	Sequence variants
Stimulus block 1	Standard "A-A-B" (90%)	LE-LE-JE; LE-LE-WE; LE-LE-DI; LE-LE-LI; WI-WI-JE; WI-WI-WE; WI-WI-DI; WI-WI-LI; JI-JI-JE; JI-JI-WE; JI-JI-DI; JI-JI-LI; DE-DE-JE; DE-DE-WE; DE-DE-DI; DE-DE-LI
	Pattern-obeying deviant "A-A-B" (5%)	BA-BA-BO, KO-KO-GE
	Pattern-violating deviant "A-B-B" (5%)	BA-PO-PO, KO-GA-GA
Stimulus block 2	Standard "A-B-B" (90%)	LE-JE-JE; LE-WE-WE; LE-DI-DI; LE-LI-LI; WI-JE-JE; WI-WE-WE; WI-DI-DI; WI-LI-LI; JI-JE-JE; JI-WE-WE; JI-DI-DI; JI-LI-LI; DE-JE-JE; DE-WE-WE; DE-DI-DI; DE-LI-LI
	Pattern-obeying deviant "A-B-B" (5%)	BA-BO-BO, KO-GE-GE
	Pattern-violating deviant "A-A-B" (5%)	BA-BA-PO, KO-KO-GA

One of the structures (AAB or ABB, in different stimulus blocks) was assigned to standards and pattern-obeying deviants. The other structure was assigned to pattern-violating deviants. Sixteen variants of standard sentences were used in both stimulus blocks to exclude the possibility of standards being memorized by the brain as individual objects. In both type of deviants, two variants were applied per stimulus block. The percentages refer to the proportion of each of the stimulus categories out of the total number of sentences (996).The stimulus block 1 was applied for one animal group (n = 7) and stimulus block 2 for the other animal group (n = 7).

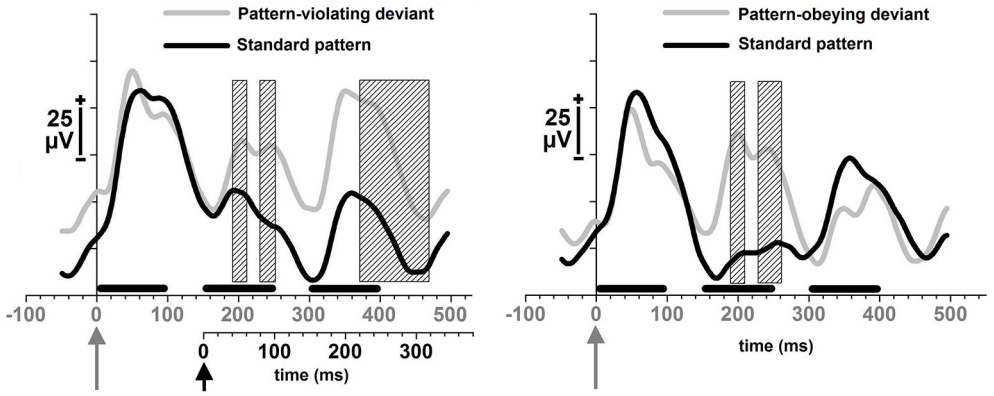


FIGURE 1 | Local-field potentials in response to pseudo-sentences. Responses to pattern-violating deviants and standard sentences immediately preceding them (left); and responses to pattern-obeying deviants and standard sentences immediately preceding them (right). The horizontal black bars represent each of the three pseudowords of 100 ms in duration. The triplets were presented at 150 ms stimulus-onset-asynchrony. The gray arrow in figures refers to the onset of the first word of a deviant sentence that physically differed from the standards; and the black arrow in the figure on the left refers to the onset of the structural change present only in the pattern-violating deviants. The two different time scales at the bottom of the left figure refer to the two different onsets of the different deviances in the pattern-violating deviant sentences (onset of the physical difference—the gray time line; onset of the pattern-related difference—the black time line). Shaded rectangles illustrate the time windows of significant amplitude differences ($p < 0.05$) between the two waveforms as indicated by point-by-point *t*-tests.

and 228–261 ms after the sentence onset, [$t_{(13)} = 2.2$–2.8, $p = 0.016$–0.048].

The second MMR was found only for the pattern-violating deviant sentences, in which the second word at a low probability (probability 0.05) violated the pattern that the rest of the sentences followed (probability 0.95). The latency for this MMR second was 217.5–316.5 ms from the onset of the second word, [$t_{(13)} = 2.2$–3.6, $p = 0.003$–0.050] (**Figure 1**, left).

Next, an ANOVA comparing the responses to the pattern-violating and pattern-obeying deviants and their consecutive standards in the time window in which the second MMR was found (i.e., 217.5–316.5 ms from the onset of the second word) was conducted. Significant interaction effect of stimulus type × deviant type was found, [$F_{(1, 13)} = 8.7$, $p = 0.011$, $\eta_p^2 = 0.401$]. Main effects were non-significant. Responses to pattern-violating deviant sequences and those to the preceding standard sequences differed significantly, [$t_{(13)} = 3.5$, $p = 0.004$, $d = 1.02$]. The corresponding difference was non-significant for the pattern-obeying deviants and preceding standards, [$t_{(13)} = 0.7$, $p = 0.525$, $d = 0.23$]. **Figure 2** depicts the mean amplitude values, standard deviation, and individual subjects' amplitude values for the differential responses.

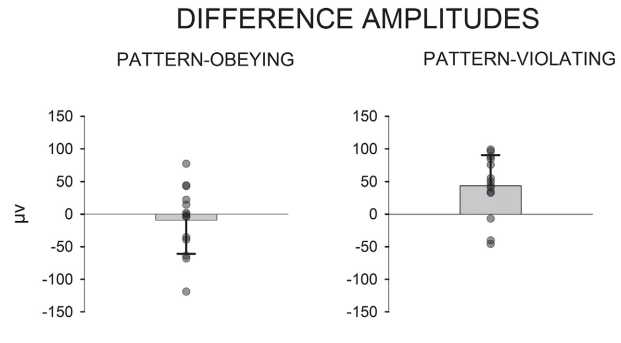


FIGURE 2 | Mean amplitude values, standard deviation and scatterplots for the individual animals' amplitude values for the second MMR (217.5–316.5 ms from the onset of the second word). Differential LFPs (deviant - standard) to pattern-obeying and pattern-violating deviant sentences.

DISCUSSION

Both types of deviant sentences, pattern-obeying and pattern-violating deviants, were detected from among the repeated standard sentences in the rat brain as indexed by the electrophysiological mismatch response. The earlier difference starting at 187.5 ms, after the sentence onset, was most probably elicited by the physical novelty of the deviant sounds; since the probability for the each standard variant was 22.5% and that of the deviant variants was 5%. An additional mismatch response, starting at 217.5 ms from the onset of the pattern change, was specifically found for the deviant sound sequences that were different in pattern structure from the frequently presented standard sequences. This finding suggests that anesthetized rats are able to extract structural patterns from speech stream that is carried out at a fast pace, and generalize this information to new items (since the deviant sentences differed physically from the standard sentences). Namely, in order to detect the pattern-violating deviant sequences, the brains of the animals needed to make a representation of the structure in the frequently presented "standard" sequences (Näätänen et al., 2001, 2010).

There is previous evidence of non-human animals' ability to extract grammatical rules from speech sounds. Common marmosets (New World monkeys) detected the grammatical differences based on simpler learning strategies than Rhesus monkeys (Old Wold monkeys) (Wilson et al., 2013). Similar ability for rule extraction has been previously reported from sinusoidal sounds in rats (Murphy et al., 2008) and from speech-specific calls in

song birds (e.g., Gentner et al., 2006). In human infants, there is evidence that they learn more easily rule-like regularities from speech than from other auditory material (Marcus et al., 2007). It is not known whether this preference is related to linguistic potential in an infant's brain, familiarity of the speech sounds, or some other factors. Future studies in non-human animals could enlighten this issue.

In the present study we tested the rats' ability to detect pattern violation is speech sound sequences that all included a repetition of an element. Therefore, they were not possible to differentiate by detecting only this property of the stimulus. On the other hand, the generalization of the present results may be restricted to stimuli in which the pattern is defined as a repetition of an element and only the position of the repetition in the three-syllabic sequence is varied. Humans are particularly sensitive to rules that are expressed as a repetition of an element at the edges of a sequence (Endress et al., 2005). In our experiment, repetitions were always at the edge of the sequence. It is thus unclear as to what extent the present results in rats can be generalized to other types of rules. Furthermore, the types of rules applied to the previous studies on rule extraction have been under debate (Gentner et al., 2010; ten Cate et al., 2010). Thus, far studies in song birds have been progressive in solving this problem (e.g., van Heijningen et al., 2013), but there are still open questions (ten Cate and Okanoya, 2012). Electrophysiological methods which provide accurate information on the timing of neural activity (recorded in animals and humans) would be a feasible addition when studying different levels of cognitive complexity required in rule extraction. In humans, event-related potentials to study processing of non-adjacent dependencies, i.e., AXC structure in which the first and the last element are dependent (De Diego Balaguer et al., 2007; Mueller et al., 2009) and structural rules (ABB vs. ABA, Sun et al., 2012) in speech sounds have been utilized.

Previous behavioral research has failed to find evidence for rule extraction from speech sounds in rats (Toro and Trobalon, 2005). In this study, rats were presented with similar three-syllabic sequences of speech sounds, as in Marcus et al. (the third experiment, 1999). Our stimuli were nearly identical and the variability in the "standard" and "deviant" sequences was also the same (16 standard variants and 2 deviant variants of both deviant types). In the study by Toro and Trobalon (2005), rats indicated the detection of the pattern violation by pressing a lever. The present positive finding may be related to the methodology used. Namely, the mismatch response is known to be capable of probing into auditory cognition regardless of its behavioral manifestations (Tremblay et al., 1998). This method can bypass a wide range of factors related to behavior, for example, motivation, attention, or requirements of overt behavior. However, the constraints of such non-behavioral measures should also be acknowledged. Namely, it is unclear whether this ability can support behavioral adaptation in rats or not. Nevertheless, its existence in an animal species, which do not use complex sequences of calls in intra-species communication, (as compared to human speech or birdsong, e.g., Doupe and Kuhl, 1999; Gentner et al., 2006) supports the notion of its non-linguistic origin. Moreover, these findings endorse the view that even the most complex functions, quintessentially considered inherent to the human brain only, may in fact, also be represented in a primitive form (Näätänen et al., 2010) in brains thus far considered evolutionarily incapable of such procedures. Since extraction of rule-like patterns, in serially presented spectro-temporally complex sounds, is one of the mechanisms utilized by humans in receptive language learning the results might imply that some of the mechanisms supporting human language learning may not have evolved solely for human language during evolution.

In conclusion, the present results demonstrate the ability of the anesthetized rat brain to detect and represent the common abstract rule or pattern obeyed by a sequence of speech-like sound stimuli with a wide acoustic variation. Hence, these results appear to give a major contribution to the evidence suggesting the presence of the automatic sensory-cognitive core of cognitive function that is shared by humans and different other, at least higher species, at different developmental stages, and even in different states of consciousness, as proposed by Näätänen et al. (2001, 2010).

AUTHOR CONTRIBUTIONS

All authors contributed substantially to the conception and design of the work; Tanel Mällo, Timo Ruusuvirta, and Piia Astikainen contributed to the acquisition and analysis of the data, and all authors contributed to the interpretation of data for the work; Tanel Mällo and Piia Astikainen drafted the work, and Timo Ruusuvirta and Risto Näätänen contributed to revising it critically for its important intellectual content. Final approval of the version to be published was attained from all authors who also agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, grant number 127595 and 273134 to Piia Astikainen and grant number 122743 to Risto Näätänen The authors are grateful to Petri Kinnunen for preparing the stimulus materials, M. Sci. Mustak Ahmed for assisting in electrophysiological recordings, and Dr. Markku Penttonen for technical help.

REFERENCES

Ahmed, M., Mällo, T., Leppänen, P. H. T., Hämäläinen, J., Äyräväinen, L., Ruusuvirta, T., et al. (2011). Mismatch brain response to speech sound changes in rats. *Front. Psychol.* 2:283. doi: 10.3389/fpsyg.2011.00283

Astikainen, P., Ruusuvirta, T., and Näätänen, R. (2014). Rapid categorization of sound objects in anesthetized rats as indexed by the electrophysiological mismatch response. *Psychophysiology* 51, 1195–1199. doi: 10.1111/psyp.12284

Astikainen, P., Ruusuvirta, T., Wikgren, J., and Penttonen, M. (2006). Memory-based detection of rare sound feature combinations in anesthetized rats. *Neuroreport* 17, 1561–1564. doi: 10.1097/01.wnr.0000233097.13032.7d

De Diego Balaguer, R., Toro, J. M., Rodriguez-Fornells, A., and Bachoud-Levi, A. C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS ONE* 2:e1175. doi: 10.1371/journal.pone.0001175

Dooling, R. J., and Brown, S. D. (1990). Speech perception by budgerigars (Melopsittacusundulatus): spoken vowels. *Percept. Psychophys.* 47, 568–574. doi: 10.3758/BF03203109

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Ann. Rev. Neurosci.* 22, 567–631. doi: 10.1146/annurev.neuro.22.1.567

Endress, A. D., Scholl, B. J., and Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *J. Exp. Psychol. Gen.* 134, 406–419. doi: 10.1037/0096-3445.134.3.406

Engineer, C. T., Perez, C. A., Chen, Y. H., Carraway, R. S., Reed, A. C., Shetake, J. A., et al. (2008). Cortical activity patterns predict speech discrimination ability. *Nat. Neurosci.* 11, 603–608. doi: 10.1038/nn.2109

Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature* 440, 1204–1207. doi: 10.1038/nature04675

Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (2010). Simple stimuli, simple strategies. *Proc. Natl. Acad. Sci. U.S.A.* 107:E65. doi: 10.1073/pnas.1000501107

Hauser, M. D., Newport, E. L., and Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition* 78, B53–B64. doi: 10.1016/S0010-0277(00)00132-3

Hauser, M. D., Weiss, D., and Marcus, G. (2002). Rule learning by cotton-top tamarins. *Cognition* 86, B15–B22. doi: 10.1016/S0010-0277(02)00139-7

Kraus, N., McGee, T., Carrell, T., King, C., Littman, T., and Nicol, T. (1994). Discrimination of speech-like contrasts in the auditory thalamus and cortex. *J. Acoust. Soc. Am.* 96, 2758–2768. doi: 10.1121/1.411282

Maggi, C. A., and Meli, A. (1986). Suitability of urethane anesthesia for physiopharmacological investigations in various systems Part 1: general considerations. *Experientia* 42, 109–114. doi: 10.1007/BF01952426

Marcus, G. F., Fernandes, K. J., and Johnson, S. J. (2007). Infant rule learning facilitated by speech. *Psychol. Sci.* 18, 387–391. doi: 10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., Vijayan, S., BandiRao, S., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science* 283, 77–80. doi: 10.1126/science.283.5398.77

Mueller, J. L., Oberecker, R., and Friederici, A. D. (2009). Syntactic learning by mere exposure - an ERP study in adult learners. *BMC Neurosci.* 10:89. doi: 10.1186/1471-2202-10-89

Murphy, R. A., Mondragon, E., and Murphy, V. A. (2008). Rule learning by rats. *Science* 319, 1849–1851. doi: 10.1126/science.1151564

Näätänen, R., Astikainen, P., Ruusuvirta, T., and Huotilainen, M. (2010). Automatic auditory intelligence: an expression of the sensory-cognitive core of cognitive processes. *Brain Res. Rev.* 64, 123–136. doi: 10.1016/j.brainresrev.2010.03.001

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

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385, 432–434. doi: 10.1038/385432a0

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., and Winkler, I. (2001). "Primitive intelligence" in the auditory cortex. *Trends Neurosci.* 24, 283–288. doi: 10.1016/S0166-2236(00)01790-2

Retraction notice. (2010). Retraction notice. Rule learning by cotton-top tamarins. Cognition 86, B15–B22. *Cognition* 117:106.

Ruusuvirta, T., Koivisto, K., Wikgren, J., and Astikainen, P. (2007). Processing of melodic contours in urethane-anaesthetized rats. *Eur. J. Neurosci.* 26, 701–703. doi: 10.1111/j.1460-9568.2007.05687.x

Sinnott, J. M., Beecher, M. D., Moody, D. B., and Stebbins, W. C. (1976). Speech sound discrimination by monkeys and humans. *J. Acoust. Soc. Am.* 60, 687–695. doi: 10.1121/1.381140

Sinnott, J. M., and Mosteller, K. W. (2001). A comparative assessment of speech sound discrimination in the *Mongolian gerbil. J. Acoust. Soc. Am.* 110, 1729–1732. doi: 10.1121/1.1398055

Sun, F., Hoshi-Shiba, R., Abla, D., and Okanoya, K. (2012). Neural correlates of abstract rule learning: an event-related potential study. *Neuropsychologia* 50, 2617–2624. doi: 10.1016/j.neuropsychologia.2012.07.013

ten Cate, C., and Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. *Philos. Trans. R. Soc. Biol. Sci.* 367, 1984–1994. doi: 10.1098/rstb.2012.0055

ten Cate, C., van Heijningen, C. A. A., and Zuidema, W. (2010). Reply to Gentner et al.: as simple as possible, but not simpler. *Proc. Natl. Acad. Sci. U.S.A.* 107, E66–E67. doi: 10.1073/pnas.1002174107

Toro, J. M., and Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. *Percept. Psychophys.* 67, 867–875. doi: 10.3758/BF03193539

Tremblay, K., Kraus, N., and McGee, T. (1998). The time course of auditory perceptual learning: neurophysiological changes during speech–sound training. *Neuroreport* 9, 3557–3560. doi: 10.1097/00001756-199811160-00003

van Heijningen, C. A., Chen, J., van Laatum, I., van der Hulst, B., and ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Anim. Cogn.* 16, 165–175. doi: 10.1007/s10071-012-0559-x

Wilson, B., Slater, H., Kikuchi, Y., Milne, A. E., Marslen-Wilson, W. D., Smith, K., et al. (2013). Auditory artificial grammar learning in macaque and marmoset monkeys. *J. Neurosci.* 33, 18825–18835. doi: 10.1523/JNEUROSCI.2414-13.2013

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: 25 June 2014; accepted: 30 October 2014; published online: 17 November 2014.

*Citation: Astikainen P, Mällo T, Ruusuvirta T and Näätänen R (2014) Electrophysiological evidence for change detection in speech sound patterns by anesthetized rats. Front. Neurosci. **8**:374. doi: 10.3389/fnins.2014.00374*

This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 August 2014
doi: 10.3389/fnins.2014.00240

Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings

Mitchell Steinschneider[1]*†, Kirill V. Nourski[2†], Ariane E. Rhone[2], Hiroto Kawasaki[2], Hiroyuki Oya[2] and Matthew A. Howard III[2]

[1] Departments of Neurology and Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA
[2] Human Brain Research Laboratory, Department of Neurosurgery, The University of Iowa, Iowa City, IA, USA

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Christopher I. Petkov, Newcastle University, UK
Dana Boatman-Reich, Johns Hopkins School of Medicine, USA

****Correspondence:***
Mitchell Steinschneider, Department of Neurology, Albert Einstein College of Medicine, 1300 Morris Park Ave., Bronx, NY, 10461, USA
e-mail: mitchell.steinschneider@einstein.yu.edu

† These authors have contributed equally to this work

Speech perception requires that sounds be transformed into speech-related objects with lexical and semantic meaning. It is unclear at what level in the auditory pathways this transformation emerges. Primary auditory cortex has been implicated in both representation of acoustic sound attributes and sound objects. While non-primary auditory cortex located on the posterolateral superior temporal gyrus (PLST) is clearly involved in acoustic-to-phonetic pre-lexical representations, it is unclear what role this region plays in auditory object formation. Additional data support the importance of prefrontal cortex in the formation of auditory objects, while other data would implicate this region in auditory object selection. To help clarify the respective roles of auditory and auditory-related cortex in the formation and selection of auditory objects, we examined high gamma activity simultaneously recorded directly from Heschl's gyrus (HG), PLST and prefrontal cortex, while subjects performed auditory semantic detection tasks. Subjects were patients undergoing evaluation for treatment of medically intractable epilepsy. We found that activity in posteromedial HG and early activity on PLST was robust to sound stimuli regardless of their context, and minimally modulated by tasks. Later activity on PLST could be strongly modulated by semantic context, but not by behavioral performance. Activity within prefrontal cortex also was related to semantic context, and did co-vary with behavior. We propose that activity in posteromedial HG and early activity on PLST primarily reflect the representation of spectrotemporal sound attributes. Later activity on PLST represents a pre-lexical processing stage and is an intermediate step in the formation of word objects. Activity in prefrontal cortex appears directly involved in word object selection. The roles of other auditory and auditory-related cortical areas in the formation of word objects remain to be explored.

Keywords: electrocorticography, Heschl's gyrus, high gamma, prefrontal cortex, semantics, speech, superior temporal gyrus

INTRODUCTION

Speech perception requires that incoming sounds be transformed into word objects. It is unclear at what level in the auditory pathways this transformation occurs. Some data suggest that primary auditory cortex principally represents acoustic sound attributes (Mesgarani et al., 2008; Poeppel et al., 2008; Steinschneider et al., 2013). Other data suggest that primary auditory cortex is more directly involved in sound object representation (Nelken, 2008; Nelken and Bar-Yosef, 2008). It is also unclear what role non-primary auditory cortex, located on the posterolateral superior temporal gyrus (PLST), plays in object formation. PLST is critical for acoustic-to-phonetic transformations (Boatman, 2004; Poeppel et al., 2008; Chang et al., 2010; Steinschneider et al., 2011; Mesgarani et al., 2014). This process could be interpreted as a remapping of the speech signal from one encoding acoustic attributes to one representing its phonemic components. By extension, it could be argued that this process remains a precursor to the formation of word objects. In this scheme, word object formation would be expected to take place at higher levels in auditory and auditory-related cortex (Griffiths and Warren, 2004; Griffiths et al., 2012).

Multiple studies have examined the transformation of neural activity associated with the representation of sound attributes to a representation of sound objects (Griffiths and Warren, 2004; Winkler et al., 2006; Shinn-Cunningham, 2008; Alain and Winkler, 2012; Griffiths et al., 2012; Simon, 2014). At the object

Abbreviations: ECoG, electrocorticography; ERBP, event-related band power; ERP, event-related potential; fMRI, functional magnetic resonance imaging; HG, Heschl's gyrus; IFG, inferior frontal gyrus; MFG, middle frontal gyrus; MRI, magnetic resonance imaging; MTG, middle frontal gyrus; PLST, posterolateral superior temporal gyrus; POA, place of articulation; STG, superior temporal gyrus; VOT, voice onset time.

formation processing stage, neural activity associated with a specific object must be distinct from that associated with other sound objects. Further, the neural representation of an object must be relatively invariant to variations in the detailed acoustics of the sounds. For instance, the representation of a specific word and its meaning must remain stable despite variations in acoustic characteristics that occur when a given word is spoken by different talkers. Given these requirements, object formation can be evaluated by utilizing tasks that require classifying words into semantic categories (Shahin et al., 2006; Hon et al., 2009).

Intracranial electrophysiological recordings in humans offer a unique opportunity for studying task-related activity in auditory cortex that accompanies semantic processing of speech. The technique combines exquisite spatial and temporal resolution beyond that offered by non-invasive methods such as neuromagnetic responses and functional magnetic resonance imaging (MRI) (e.g., Lachaux et al., 2012). An excellent example of the sensitivity and specificity provided by intracranial recordings in humans is the study demonstrating that competing speech signals can be segregated according to speaker through analysis of cortical activity recorded from PLST during selective attention tasks (Mesgarani and Chang, 2012). The neural activity associated with the attended stream was enhanced, while activity associated with the unattended stream was suppressed. In a related study, target detection tasks led to enhanced neural activity to target tone stimuli on PLST when compared to responses obtained during passive listening and responses to non-target tone stimuli (Nourski et al., 2014a). These effects occurred during later portions of the neural responses. Early activity was minimally affected by the task requirement and appeared to represent the acoustic attributes of the tones. Similarly, minimal effects were noted in activity simultaneously recorded from posteromedial Heschl's gyrus (HG), the putative location of core auditory cortex. These findings suggest that activity generated within posteromedial HG and early activity from PLST reflect acoustic encoding rather than the representation of non-speech and speech-related objects at the phonemic level. It remains unclear from these studies, however, if this region of auditory cortex will also be involved in the formation of speech-related objects at the level of words and their semantic meaning.

The current study focused on high gamma responses (70–150 Hz) generated during target detection tasks using both speech and non-speech stimuli. High gamma activity has been shown to be a sensitive and specific indicator of auditory cortical activation and has been successfully used to define organizational features of human auditory cortex (e.g., Crone et al., 2001; Steinschneider et al., 2008, 2011; Flinker et al., 2010; Mesgarani and Chang, 2012; Mesgarani et al., 2014; Nourski et al., 2014b). Tasks of the current study included detecting words belonging to specific semantic categories or talker gender, as well as the detection of tones intermixed with the word sequences. Words were consonant-vowel-consonant exemplars from the semantic categories of animals, numbers and colors, as well as nonsense syllables, each spoken by different male and female talkers. Therefore, neural activity associated with target detection should not be based solely on acoustic attributes and instead should be related to semantic categorization and, consequently, word object formation. We predicted that the tone detection task would not engage speech-related object formation, as this task only required differentiating the sound objects based on their acoustic attributes. In contrast, tasks that required the subject to detect words from a specific target category necessitated that words be decoded and categorized as sound objects belonging to specific semantic categories. Detection of talker gender provided an intermediate control condition. If the successful completion of the task was solely dependent upon decoding the fundamental frequencies typically encountered across gender (e.g., Hillenbrand et al., 1995), then, we hypothesized, sound object formation would not engage word-specific processing. If, however, formation of word objects incorporated representation of gender, then response profiles should be similar to that observed when words were categorized along semantic categories.

We also examined neural activity within auditory-related cortical areas that have been shown to be critical components of the neural network subserving speech perception (e.g., Rauschecker and Scott, 2009). Neural activity from inferior frontal gyrus (IFG) in the language-dominant hemisphere measured with intracranial recordings has been shown to represent lexical, grammatical and phonological aspects of speech (e.g., Sahin et al., 2009). In the present study, responses from the portion of IFG that overlaps with classically defined Broca's area were compared with activity recorded from HG and PLST. Additionally, contributions from middle temporal gyrus (MTG) and middle frontal gyrus (MFG) were examined, as these higher-order cortical regions may also be involved in word object formation (Griffiths and Warren, 2004; Poeppel et al., 2008). Simultaneous recordings from multiple regions including core, non-core and auditory-related cortex provided a unique opportunity to examine the role of each of these areas in word object formation during target detection tasks with high temporal and spatial detail.

METHODS

SUBJECTS

Experimental subjects were three neurosurgical patients diagnosed with medically refractory epilepsy and undergoing chronic invasive electrocorticographic (ECoG) monitoring to identify potentially resectable seizure foci. The subjects were 38 (L258), 30 (L275), and 40 (L282) years old. All subjects were male, right-handed and left hemisphere language-dominant, as determined by intracarotid amytal (Wada) test results. Recordings were obtained from the left hemisphere in all three subjects. Research protocols were approved by the University of Iowa Institutional Review Board and the National Institutes of Health. Written informed consent was obtained from all subjects. Research participation did not interfere with acquisition of clinically required data, and subjects could rescind consent at any time without interrupting their clinical evaluation.

All subjects underwent audiometric evaluation before the study, and none was found to have hearing deficits that should impact the findings presented in this study. Subjects L258 and L282 were native English speakers, and subject L275 was a native Bosnian speaker who learned German at the age of 10 and English at the age of 17. Neuropsychological testing of L258 was normal except for mild deficiencies in verbal working memory. Subject

L275 had grossly intact conversational language comprehension, though formal neuropsychological testing showed non-localizing cognitive function deficits. Subject L282 had 13 years earlier undergone anterior temporal lobectomy that spared auditory cortex on the superior temporal gyrus. This subject was found to have mild deficits in verbal memory, fluency and naming. However, all three subjects had comparable performance in all experimental tasks both in terms of target detection accuracy and reaction times. This indicates that their performance of the tasks was not limited by any cognitive deficits identified during formal neuropsychological testing. Intracranial recordings revealed that auditory cortical areas were not epileptic foci in any subject.

Experiments were carried out in a dedicated electrically-shielded suite in The University of Iowa General Clinical Research Center. The room was quiet, with lights dimmed. Subjects were awake and reclining in an armchair.

STIMULI

Experimental stimuli were consonant-vowel-consonant syllables [cat], [dog], [five], [ten], [red], [white], [res], and [tem] from TIMIT (Garofolo et al., 1993) and LibriVox (http://librivox.org/) databases. Non-word syllables were excised from words using SoundForge 4.5 (Sonic Foundry Inc., Madison, WI). A total of 20 unique exemplars of each syllable were used in each experiment: 14 spoken by different male and 6 by different female speakers. Additionally, the stimulus set included complex tones with fundamental frequencies of 125 (28 trials) and 250 Hz (12 trials), approximating the average voice fundamental frequencies of male and female speakers, respectively. All stimuli were normalized to the same root-mean-square amplitude and edited to be 300 ms in duration using SoundForge with 5 ms rise-fall times. They were presented with an inter-stimulus interval chosen randomly within a Gaussian distribution (mean interval 2 s; $SD = 10$ ms) to reduce heterodyning in the recordings secondary to power line noise. Stimuli were delivered via insert earphones (ER4B, Etymotic Research, Elk Grove Village, IL) that were integrated into custom-fit earmolds. Stimulus delivery was controlled using Presentation software (Version 16.5 Neurobehavioral Systems, http://www.neurobs.com/).

The same stimuli were presented in random order in multiple target detection tasks. The target stimuli were either complex tones (presented as first block in each subject), speech stimuli spoken by female talkers, or words belonging to specific semantic categories such as animals or numbers. The subjects were instructed to use the index finger of their left hand (ipsilateral to the recording hemisphere) to push the response button whenever they heard a target sound. Prior to data collection, the subjects were presented with a random-sequence preview of stimuli to ensure that the sounds were presented at a comfortable level and that they understood the task requirements.

RECORDINGS

ECoG recordings were simultaneously made from HG and lateral cortical surface using multicontact depth and subdural grid electrodes, respectively. Details of electrode implantation have been described previously, and more comprehensive details regarding recording, extraction and analysis of high gamma cortical activity are available for the interested reader (Howard et al., 1996, 2000; Reddy et al., 2010; Nourski et al., 2013; Nourski and Howard, 2014). In brief, hybrid depth electrode arrays were implanted stereotactically into HG, along its anterolateral to posteromedial axis. In subject L258, a hybrid depth electrode was used, which contained 4 cylindrical platinum macro-contacts, spaced 10 mm apart, and 14 platinum micro-contacts, distributed at 2–4 mm intervals between the macro contacts. In subjects L275 and L282, a depth electrode with 8 macro-contacts, spaced 5 mm apart, was used. Subdural grid arrays were implanted over the lateral surface of temporal and frontal lobes in subjects L258 and L275. The grid arrays consisted of platinum-iridium disc electrodes (2.3 mm exposed diameter, 5 mm center-to-center inter-electrode distance) embedded in a silicon membrane. The electrodes were arranged in an 8×12 grid, yielding a 3.5×5.5 cm array of 96 contacts. A subgaleal contact was used as a reference. Electrode arrays were placed solely on the basis of clinical requirements, and were part of a more extensive set of recording arrays meant to identify seizure foci. Electrodes remained in place under the direction of the patients' treating neurologists.

Subjects underwent whole-brain high-resolution T1-weighted structural MRI scans (resolution 0.78×0.78 mm, slice thickness 1.0 mm) before electrode implantation to locate recording contacts. Two volumes were averaged to improve the signal-to-noise ratio of the MRI data sets and minimize the effects of movement artifact on image quality. Pre-implantation MRIs and post-implantation thin-sliced volumetric computerized tomography scans (resolution 0.51×0.51 mm, slice thickness 1.0 mm) were co-registered using a linear co-registration algorithm with six degrees of freedom (Jenkinson et al., 2002). Locations of recording sites were confirmed by co-registration of pre- and post-implantation structural imaging and aided by intraoperative photographs.

Data acquisition was controlled by a TDT RZ2 real-time processor (Tucker–Davis Technologies, Alachua, FL). Collected ECoG data were amplified, filtered (0.7–800 Hz bandpass, 12 dB/octave rolloff), digitized at a sampling rate of 2034.5 Hz, and stored for subsequent offline analysis. Behavioral responses to the target stimuli were recorded using a Microsoft SideWinder game controller. The timing of the button-press events was recorded and stored for analysis along with ECoG data.

DATA ANALYSIS

ECoG data obtained from each recording site were downsampled to a rate of 1000 Hz. To minimize contamination with power line noise, ECoG waveforms were de-noised using an adaptive notch filtering procedure (Nourski et al., 2013). Prior to calculation of high gamma event-related band power (ERBP), individual trials were screened for possible contamination from electrical interference, epileptiform spikes, high amplitude slow wave activity, or movement artifacts. To that end, individual trial waveforms with voltage exceeding 2.5 standard deviations from the mean were rejected from further analysis. Data analysis was performed using custom software written in MATLAB Version 7.14 programming environment (MathWorks, Natick, MA, USA).

Quantitative analysis of the ERBP focused on the high gamma ECoG frequency band. High gamma ERBP was calculated for

each recording site. Single-trial ECoG waveforms were bandpass filtered between 70 and 150 Hz (100th order finite impulse response filter) and squared. The resultant high gamma power waveforms were smoothed using a moving average filter with a span of 25 ms, log-transformed, normalized to power in a pre-stimulus reference (250-50 ms prior to stimulus onset), and averaged across trials. To assess the presence and timing of task-related modulation of high gamma activity on representative cortical sites, single-trial high gamma ERBP was first averaged in 50 ms-wide consecutive windows to decrease the number of multiple comparisons. Next, for each window from 0–50 to 950–1000 ms, a two-sample one-tailed *t*-test was performed on single-trial windowed ERBP values to compare responses to stimuli presented in the non-target (tones task) and target condition. Finally, *p*-values were corrected for multiple comparisons (i.e., recording sites and time windows) using false discovery rate by controlling the false discovery rate following the method of Benjamini and Hochberg (1995) and Benjamini et al. (2001) with a threshold of $q = 0.01$.

RESULTS

HG

Neural activity on HG primarily represented acoustic attributes of the speech stimuli (**Figure 1**). **Figure 1A** illustrates the location of the eight recording contacts that targeted HG along its long axis in subject L275. Mean high gamma power elicited by three acoustic attributes of speech is shown for each recording site (**Figure 1B**). Responses to the speech stimuli spoken by male talkers were consistently larger compared to those elicited by female talkers (**Figure 1B**, left column), reflecting differences in their fundamental frequency (male talkers: mean 125 Hz, *SD* 25 Hz; female talkers: mean 202 Hz, *SD* 36 Hz). These differences represent a contribution in the high gamma responses of phase locking to the lower fundamental frequency of the male talkers within posteromedial HG [sites (a) through (d)] (cf. Nourski and Brugge, 2011; Steinschneider et al., 2013).

Voice onset time of the initial stop consonants was also differentially represented in the high gamma activity. In general, high gamma activity peaked earlier for initial consonants with short voice onset times (VOTs) (i.e., [dog]) relative to those with more prolonged VOTs (i.e., [ten]). This effect was maximal in more central portions of HG compared to the observed effect of pitch on neural activity [sites (e), (f); **Figure 1B**, middle column]. Differences based upon initial consonant place of articulation (POA) were more subtle, likely due to the overlap in spectral content across the stimulus exemplars (e.g., site (d); **Figure 1B**, right column). These patterns of activity within HG were also observed in the other two subjects (Supplementary Figures 1, 2).

Whereas activity along most of HG was strongly modulated by the acoustic attributes of the sounds, responses in the high gamma range were only weakly affected by the target detection tasks (**Figure 2**). The left column in **Figure 2** compares neural activity to the same set of stimuli (female voices) in three blocks: when they were targets, when they were non-targets in the tone detection block, or when they were non-targets in a semantic task (numbers). A low-amplitude increase in high gamma was seen beginning within 600–650 ms after stimulus onset when female voices were the targets [site (a)], overlapping in time

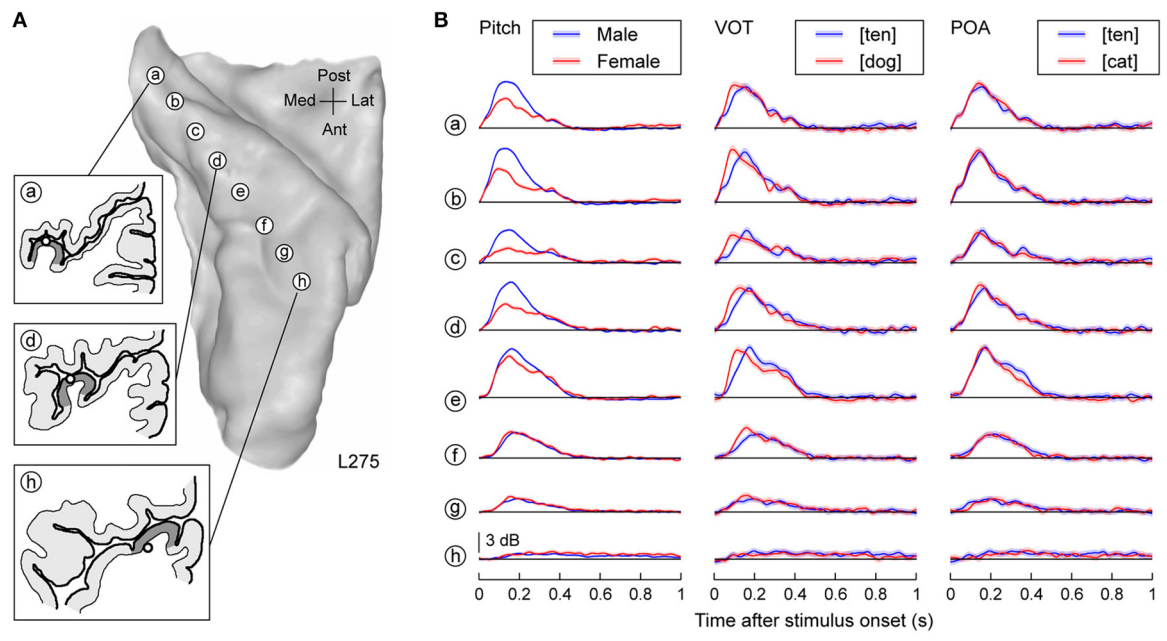


FIGURE 1 | Representation of acoustic stimulus attributes in HG. (A) MRI of left superior temporal plane in subject L275 showing the locations of recording contacts chronically implanted in HG. Insets: tracings of MRI cross-sections showing the location of three recording contacts (circles) relative to the gray matter of the HG (dark gray shading). **(B)** High gamma responses to speech sounds differing in pitch, initial stop consonant VOT and POA are shown in the left, middle and right column, respectively. Lines and shaded areas represent mean high gamma ERBP and its standard error, respectively.

with the subject's behavioral response. A similar effect was seen for responses to the animals and numbers when they were the targets. However, the onset of the task-related high gamma modulation in these semantic categorization conditions was even slower than that occurring during voice identification task ($q < 0.01$ at 750–800 ms after stimulus onset; middle and right columns of **Figure 2**).

A different pattern was observed within the most anterolateral portion of HG outside of presumed core cortex [site (h) in **Figure 2**]. Here the response was delayed relative to the activity on posteromedial HG and was specifically associated with target stimuli. Importantly, this task-related activity preceded task-related changes that were observed on posteromedial HG. These task-related increases, however, were variable across subjects. In the other two subjects, no significant task-related effects were observed at the level of either posteromedial or anterolateral HG (Supplementary Figures 3, 4). Thus, in total, task-related changes in HG were, as we will show, modest, when compared to those changes observed on PLST and in auditory-related cortex.

PLST

More complex response profiles were observed on PLST (**Figures 3**, **4**) when compared with profiles simultaneously recorded from HG (see **Figure 2**). There was a rapid and large increase in high gamma ERBP occurring within 200 ms after stimulus onset. This early activity was variably affected by the task [e.g. sites (a), (b), and (c) in **Figures 3**, **4**]. When female voices were targets, a modest but significant increase in high gamma power was observed as early as 50–100 ms after stimulus onset. Peak activity at 150–200 ms was only marginally affected by the task. Later activity was more variable across recording sites. Both enhancement of high gamma activity to the target syllables beginning prior to their offsets [e.g. sites (a), (b), and (c) in **Figure 3**] and minimal modulation of later activity related to the task (see **Figure 3B**) were observed in this region. Task-related high gamma activity was earlier than that occurring in HG (cf. **Figure 2**) and preceded the subject's behavioral response.

Responses to non-target words were also modulated by the specific task requirements. For instance, late high gamma activity to non-target words spoken by females was enhanced when the target detection tasks required words to be categorized relative to the task where complex tones were the targets (see **Figure 3**, green and blue plots, respectively). Responses to female voices when they were target stimuli were consistently larger than when they were non-targets, even though the subject was engaged in cognitively more demanding tasks (detecting numbers or animals) (see **Figure 3**, red and green plots, respectively). The difference in task difficulty can be inferred from behavioral response times, which were significantly shorter when the target was female voices (median response time 672 ms) relative to either task requiring semantic classification (animals: median response time 866 ms; numbers: median response time 815 ms) ($p < 0.001$, Mann-Whitney rank-sum tests).

Enhancement of high gamma power was also observed when the targets were animals (**Figure 4**). Once again, targets elicited the largest responses when compared to when they were non-targets presented in a tone detection task [see **Figure 4**, sites (a) and (c)]. While variable across sites [cf. site (b) in **Figure 4**], enhanced activity could occur early and remain elevated even during the time period of the behavioral response. Responses to non-target animal words presented in a different semantic categorization task (detecting numbers) were intermediate in magnitude. The behavioral reaction times were comparable in the animals and numbers detection tasks ($p = 0.71$, Mann–Whitney rank-sum test). Therefore, it is reasonable to conclude that these differences between target and relevant non-target were not based solely on task difficulty. Importantly, increases in high gamma activity observed during either semantic categorization task began prior to the offset of the syllables, suggesting that these increases were not directly related to word classification, and likely were reflecting lower-level phonological processing, a prerequisite for semantic classification (cf. Boatman, 2004).

In subject L258, task-related enhancement was not observed from sites located on PLST [Supplementary Figure 5; sites (a), (b), and (c)]. This negative finding may reflect in part differences in placement of the electrode grids, where the anterior limit of the temporal recording grid was anatomically more posterior than that in subject L275. Additionally, responses from L275 were averaged over a larger number of trials, improving signal-to-noise ratio, and subject L275 was generally more enthusiastic about performing the behavioral tasks compared to L258. However, responses were modulated by the task on sites overlying the MTG [e.g. sites (d) and (e); Supplementary Figure 5], similar to that seen on PLST in subject L275. Specifically, late responses to target stimuli were larger than responses in the tone detection task, reaching significance on site (e) in the gender identification task ($q < 0.01$), and were marginally significant ($q < 0.05$) in the semantic categorization tasks on sites (d) and (e) (significance bars are not shown). Additionally, there was a trend for non-target words to elicit larger late responses during semantic categorization tasks compared to tone detection (green and blue plots, respectively, in Supplementary Figure 5).

AUDITORY-RELATED CORTEX: IFG AND MFG

Task-related changes in high gamma activity were not restricted to the temporal lobe and were observed in IFG and MFG in both subjects with frontal lobe electrode coverage (**Figure 5**, Supplementary Figure 5). Targets elicited larger responses compared to when the same words were presented in a tone detection task in both IFG and MFG (purple bars in **Figure 5**). Minimal activity in both regions was observed in response to non-target speech stimuli when tones were targets, and phonemic and semantic processing were not necessary for task perfomance. In contrast, both targets and non-targets relevant to the task elicited responses in IFG in both subjects and MFG in subject L258 (red and green plots). Responses within MFG in subject L275 were restricted to target stimuli and had onset latencies longer than those observed at sites overlying either the superior temporal gyrus (STG), MTG or IFG, but were comparable to the timing of the late high gamma increases seen on posteromedial HG. These late increases in high gamma activity always preceded the subjects' behavioral responses (horizontal box plots in **Figure 5** and Supplementary Figure 5), which elicited high gamma activity within both pre- and post-central gyrus (see **Figures 3**, **4**).

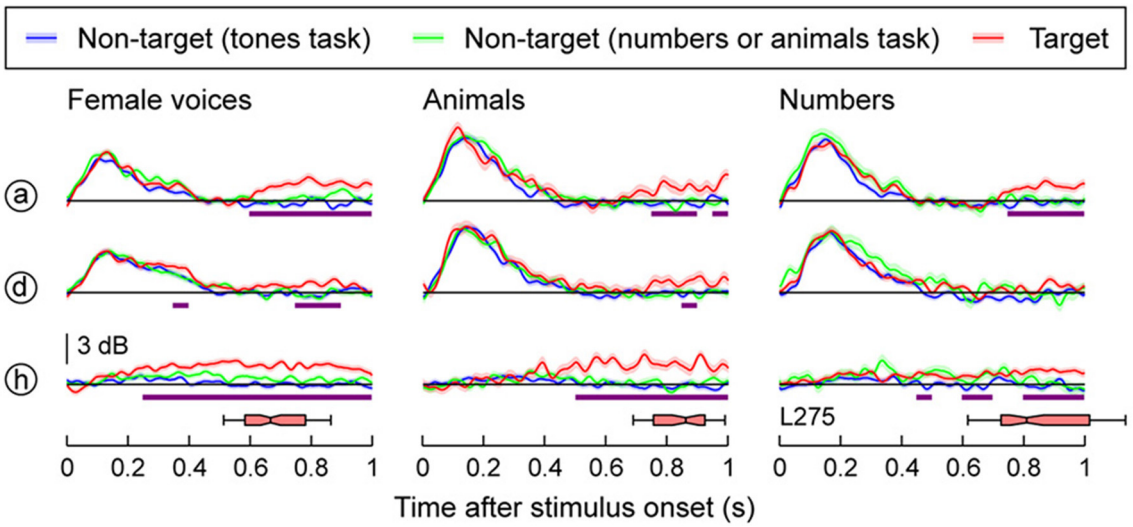


FIGURE 2 | Task effects on responses to speech stimuli in HG. Responses to three types of stimuli (female voices, animals, numbers; left, middle and right column, respectively) are shown for three representative recording sites in HG (rows). See **Figure 1A** for location of the recording sites. Colors (blue, green, and red) represent different task conditions. Lines and shaded areas represent mean high gamma ERBP and its standard error, respectively. Purple bars denote time windows where responses to the target stimuli were significantly larger than those to the same stimuli in the tones task ($q < 0.01$). Horizontal box plots denote the timing of behavioral responses to the target stimuli (medians, 10th, 25th, 75th, and 90th percentiles).

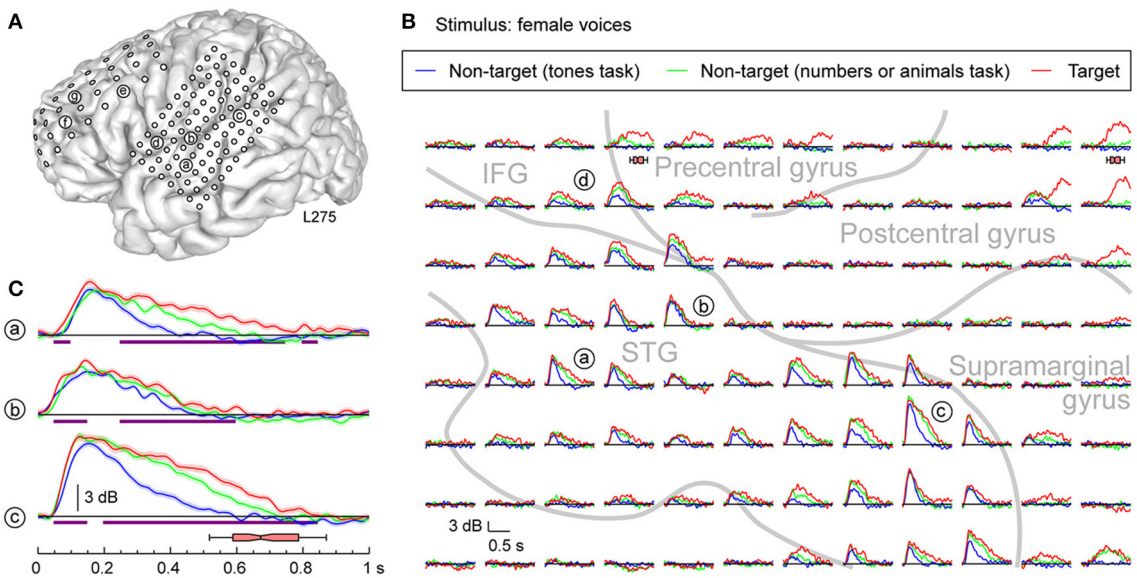


FIGURE 3 | Task effects on responses to speech stimuli (female voices) in PLST. (A) MRI of the left hemisphere in subject L275 showing the locations of chronically implanted subdural grid contacts. **(B)** High gamma responses to syllables spoken by females, presented in three different tasks (different colors), are shown for the 96-contact recording grid implanted over perisylvian cortex. Gray lines represent approximate boundaries of STG, IFG, pre- and post-central gyri covered by the recording grid. **(C)** High gamma ERBP time course replotted for three recording sites on PLST. Lines and shaded areas represent mean high gamma ERBP and its standard error, respectively. Purple bars denote time windows where responses to the target stimuli were significantly larger than those to the same stimuli in the tones task ($q < 0.01$). Horizontal box plot denotes the timing of behavioral responses to the target stimulus (median, 10th, 25th, 75th, and 90th percentiles).

DIFFERENTIAL RESPONSE PATTERNS TO TARGET STIMULI: PLST vs. IFG

Different response patterns elicited by target stimuli were noted between activity simultaneously recorded from PLST and IFG in subject L275. High gamma activity on PLST elicited by target stimuli (animals) did not significantly vary as a function of whether the subject responded rapidly or slowly or when the target was missed altogether (**Figure 6**, left column). In comparison, the same words when they were not relevant

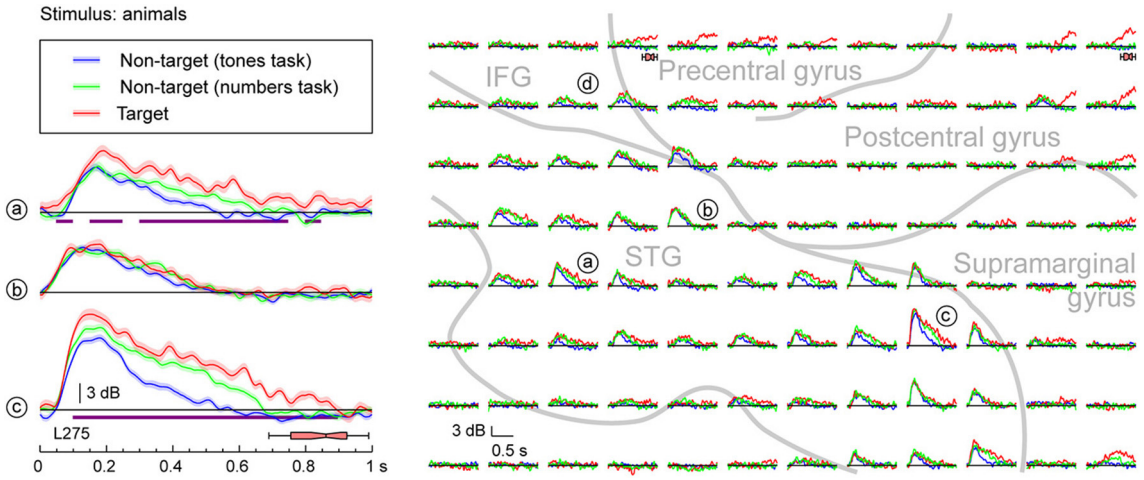


FIGURE 4 | Task effects on responses to speech stimuli (animals) in PLST. High gamma responses to speech sounds representing animals, presented in three different tasks (different colors), are shown for three representative contacts and the entire 96-contact recording grid implanted over perisylvian cortex (left and right panels, respectively). See legend of **Figure 3** for details and location of recording sites.

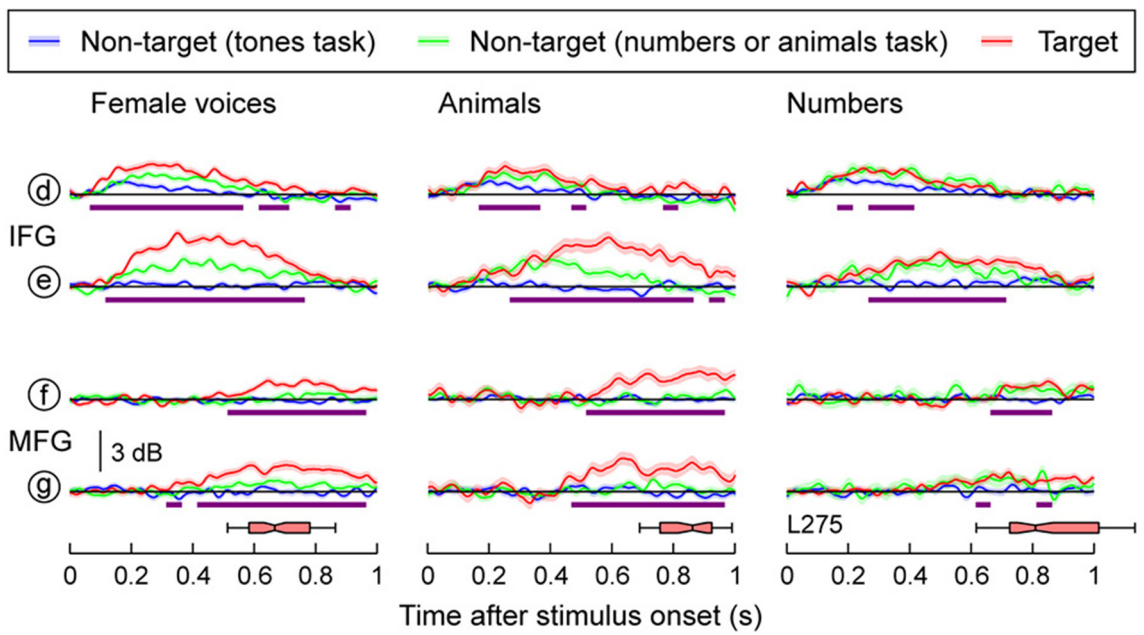


FIGURE 5 | Task effects on responses to speech stimuli in IFG and MFG. High gamma responses to three types of stimuli (female voices, animals, numbers; columns), presented in three different tasks (blue, green and red plots), are shown for two recording sites on the IFG and two sites on MFG. See **Figure 3A** for location of the sites. Lines and shaded areas represent mean high gamma ERBP and its standard error, respectively. Purple bars denote time windows where responses to the target stimuli were significantly larger than those to the same stimuli in the tones task ($q < 0.01$). Horizontal box plots denote the timing of behavioral responses to the target stimuli (medians, 10th, 25th, 75th, and 90th percentiles).

non-targets (tone detection task) elicited comparable early activity, but markedly diminished responses later in time [sites (a) and (c) in **Figure 6**]. In contrast to activity on PLST, activity within pars opercularis of IFG could be significantly modulated by the presence and timing of the behavioral response. This finding is exemplified at site (e) located on the dorsal portion of the pars opercularis (see **Figure 6**), where faster response times were associated with earlier peaks of activity when contrasted with slower behavioral responses. Additionally, misses were associated with markedly decreased responses compared to hits, and there was no response when the same stimulus was presented as a non-relevant, non-target during a tone detection task. For subject L258, parcelation of single-trial high gamma activity based on behavioral performance did not reveal consistent

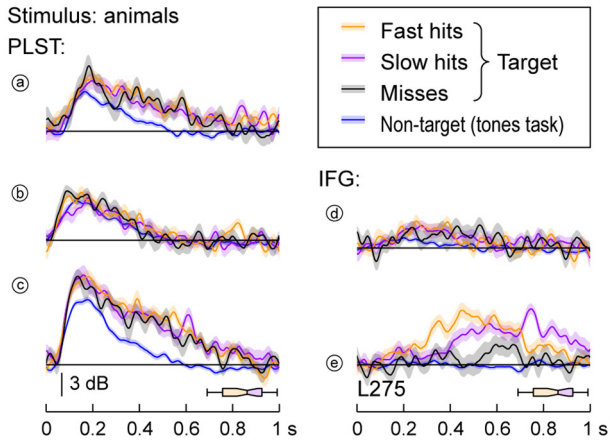


FIGURE 6 | High gamma responses to speech stimuli and the subject's behavioral performance. Data recorded from sites (a to e) (see **Figure 3A**) in response to target animal stimuli are plotted separately for trials that were associated with fast behavioral responses (17 trials; orange), slow responses (17 trials; purple), misses (11 trials; black), and non-target trials from a tone detection task (200 trials; blue). Lines and shaded areas represent mean high gamma ERBP and its standard error. Horizontal box plots denote the timing of behavioral responses to the target stimuli (medians, 10th, 25th, 75th, and 90th percentiles). See **Figure 3A** for location of the sites.

differences between PLST and IFG. This was due to highly variable responses and low response magnitudes, particularly in IFG.

DISCUSSION

POSTEROMEDIAL HG

As expected from previous studies, activity within posteromedial HG was highly sensitive to the acoustic characteristics of speech (e.g., Nourski et al., 2009; Steinschneider et al., 2013). In general, high gamma activity was greater for male talkers than female talkers. This finding reflects contribution from phase-locked responses to the lower fundamental frequency of male talkers relative to female talkers and was most prominently observed in the most posteromedial aspect of HG. This property is not unique to speech, as this region exhibits reliable phase-locked responses elicited by click trains at repetition rates of up to 200 Hz (Brugge et al., 2008, 2009; Nourski and Brugge, 2011). VOT was reflected in the timing of high gamma activity as a delay in the peak of high gamma response. This effect was most prominent in more central areas of HG, contrasting with the temporal representation of the voice fundamental. This apparent spatial differentiation may be a consequence of the tonotopic organization, wherein phase locking would most likely occur in high best frequency areas of the HG, whereas VOT would be represented in low frequency regions, due to the later onset of low frequency energy associated with voicing onset (Steinschneider et al., 1994). The absence of single and double-on responses previously reported (e.g., Steinschneider et al., 2013) can be attributed to the temporal smearing inherent to averaging of responses to unique and naturally-elicited speech exemplars characterized by different VOTs. Finally, responses reflecting differences in stop consonant POA were more subtle, and were likely a result of spectral smearing due to averaging of responses to 20 different exemplars of [cat] and [ten] across multiple talkers and the location of the recording sites with reference to the tonotopic organization of HG.

Activity within posteromedial and central HG was not strongly modulated by task requirements in all three subjects, and if it occurred (e.g., L275), it was later than task-related modulations in all other regions studied. Thus, current findings do not support the premise that human primary auditory cortex is the location where auditory object formation occurs. In contrast, studies in primary auditory cortex of experimental animals have shown robust responses reflecting auditory object formation, task-related activity, and reward expectancy (e.g., Fritz et al., 2003; Nelken and Bar-Yosef, 2008; Brosch et al., 2011; Niwa et al., 2012). The difference between the current observations and those in animals may reflect species differences and the relative complexity of auditory cortical organization in humans (Hackett, 2007). This complexity would be paralleled by greater functional specialization for primary and non-primary areas as the demands for vocal learning and auditory sequence learning become progressively more complex (Petkov and Jarvis, 2012).

Our findings in HG are consistent with several magnetoencephalograpgy and event-related potential (ERP) studies (Shahin et al., 2006; Gutschalk et al., 2008; Sabri et al., 2013; Simon, 2014; but see Bidet-Caulet et al., 2007). One study observed that during selective attention to one speech stream over another, the M100, but not M50 component of the neuromagnetic response, was modulated by the attended stream (Simon, 2014). This finding is consistent with our negative results, as the M50 component is dominated by generators in or near primary auditory cortex, while the M100 component reflects generators from multiple non-primary areas, particularly those in planum temporale (Liégeois-Chauvel et al., 1994). Another study sorted magnetoencephalograpy data according to whether or not target tones were detected in a multi-tone cloud background capable of producing informational masking of the targets (Gutschalk et al., 2008). Detected targets elicited an M100-like component that was not present when the target sounds were not detected. In contrast, both detected and undetected tones evoked auditory middle-latency and steady-state responses whose generators likely include prominent contributions from the primary auditory cortex on HG. It should be noted, however, that other studies utilizing auditory detection paradigms failed to find modulation of the N100 component (Shahin et al., 2006; Sabri et al., 2013). This negative result is not restricted to the auditory modality and has been observed in early cortical activity during visual target detection tasks (Bansal et al., 2014).

The minimal modulation of early high gamma activity that we observed replicates the findings in a previous intracranial study, where no effect was observed in the magnitude or timing of high gamma activity within posteromedial HG during a tone detection task relative to passive listening (Nourski et al., 2014a). Finally, functional neuroimaging studies have not shown consistent task-related changes in HG (Pugh et al., 1996; Leicht et al., 2010). When present, attention-related modulations occurred mainly in non-primary auditory cortex lateral to core areas (Petkov et al.,

2004). This latter finding is consistent with task-related modulations currently seen in the most anterolateral portion of HG in one subject (see **Figure 2**). It must be acknowledged, however, that limited sampling inherent to human HG recordings may be responsible for the lack of consistent task-related effects seen in the three subjects studied here.

PLST

Early activity on PLST, occurring within 200 ms after stimulus onset, was not strongly modulated by task requirements, mirroring a result seen in different subjects performing a tone detection task (Nourski et al., 2014a). Studies have demonstrated that early high gamma activity reflects more automatic processing that helps represent specific spectral characteristics of tone stimuli (Nourski et al., 2014a,b), as well as the remapping of acoustic speech characteristics to those representing phonetic categories (Chang et al., 2010; Travis et al., 2013; Mesgarani et al., 2014). In contrast, later high gamma activity on PLST could be strongly modulated by task requirements. Findings such as these are neither unique to humans nor restricted to the auditory system. For instance, during visual object detection tasks, single unit activity from neurons within areas V4 and IT of the monkey showed limited modulation as a function of the target stimulus in the initial response component, yet were strongly dependent on the specific target in later response segments (Chelazzi et al., 1998, 2001). The authors suggested that these later effects were based on feedback from higher visual centers involved in working memory, and reflected response bias toward the behaviorally relevant objects. A similar "top-down" mechanism that biases responses toward task-relevant stimuli may also be responsible for the currently observed effects in PLST.

Several studies have shown that neural patterns of activity in auditory cortex independently encode speaker identity and phonemic content of verbal speech ("Who" is saying "what"; e.g., Formisano et al., 2008; Mesgarani and Chang, 2012). We examined whether similar patterns independently encoding voice vs. speech content would emerge during the performance of the current target detection tasks, but found no clear differences. It should be noted, however, that in the study of Formisano et al. (2008), subjects passively listened to only three vowels spoken by three talkers. Here, subjects actively listened to 180 unique word exemplars spoken by an almost equal number of different talkers presented during semantic classification tasks and control conditions that included gender identification. Furthermore, the brain regions associated with gender identification were primarily located over the non-dominant right hemishere and distributed on the lateral portion of HG and Heschl's sulcus, as well as portions of the superior temporal sulcus (Formisano et al., 2008). The current study examined the dominant left hemisphere with limited sampling of HG, and did not sample neural activity in Heschl's sulsus or the superior temporal sulcus. In the study by Mesgarani and Chang (2012), the subjects were performing a different behavioral task (selective attention), and the neural activity only had to be capable of discriminating sentences spoken by two talkers (one male). It thus remains to be determined whether high gamma power, at least within PLST, is capable of independently determining multiple speaker identities (or gender) and phonemic content (e.g., Obleser and Eisner, 2009).

Response enhancement on PLST began prior to word offset during the semantic classification tasks (see **Figure 6**). The timing of response enhancement indicates that the effect was not driven by processes directly reflecting semantic classification, but instead represented the phonemic processing that must by necessity occur earlier in order to accurately decode the words. Further, the target words elicited a larger response than non-target words. As pointed out by Hon et al. (2009), any target enhancement that occurs within early sensory regions when a semantic target is detected must originate from higher-level brain areas providing relevant feedback to the lower areas. In the present study, subjects had been primed to know that the same two exemplar words for each semantic category would be presented in each successive recording block. This priming would allow subjects to know that, for instance, in the animals task, /d/ and /k/ would be the first phonemes in the target words ([dog] and [cat]) and thus provide additional information useful for the completion of the semantic task.

Response enhancement on PLST was also independent of task performance accuracy and reaction time. The same effect has been observed on PLST in a different subject performing a tone detection task, thus replicating current findings (Nourski et al., 2014a). Object-based detection tasks require two sequential processes, object formation followed by object selection (Shinn-Cunningham, 2008). The independence of the neural responses from behavioral measures are consistent with PLST being involved in the process of semantic object formation, yet not directly tied to the process of object selection. Similar observations have been made in the lateral belt field AL in macaque auditory cortex when performing a discrimination task using consonant-vowel syllables (Tsunada et al., 2011). In that study, single-cell responses reflected the categorization of the syllable (i.e., object formation), but did not vary as a function of the animal's behavioral performance (i.e., object selection). Activity that does not vary with behavioral performance likely reflects processes that precede sound object formation.

Even in the subject where later activity was strongly modulated (L275), effects were not uniform and showed site-by-site variability. This variability may partly explain why task-related modulation on PLST was not seen in subject L258 (see Supplementary Figure 5). Additionally, electrode array placement was more posterior along the STG in subject L258 when compared to the placement in L275. Electrical stimulation in subjects with epilepsy while they participated in various auditory and speech-related tasks has demonstrated the functional heterogeneity of the STG (Boatman, 2004), indicating that differences in electrode placement can be a major source of inter-subject variability. Finally, language processing skills of the subjects and effort necessary for successful performance of the task, may have also been a significant factor contributing to the inter-subject variability observed in this study.

AUDITORY-RELATED CORTEX

Multiple brain regions outside of the classically defined auditory cortex were differentially activated during the target detection

tasks. For instance, task-related activity was shown within MTG, and enhancement of later activity was observed in responses to targets and non-targets in the semantic categorization tasks. Similar activation of MTG immediately adjacent to the superior temporal sulcus in response to speech has been reported (**Figure 3** in Flinker et al., 2010). This region has been shown to be important in lexical processing, and is activated even during passive presentation of words (Dronkers et al., 2004; Indefrey and Cutler, 2004; Hagoort, 2005; Hickok, 2009). Unfortunately, sampling was too limited to better describe these modulations outside of observing that they had latencies comparable with those seen on PLST and frontal regions.

The IFG of the dominant left hemisphere was also activated during target detection tasks. High gamma activity was observed when stimuli were targets, and, to a lesser degree, non-targets. Findings are in keeping with other auditory target detection studies. Bilateral activation of the IFG occurred during an auditory detection task using positron emission tomography when targets were words, consonant-vowel syllables, or tone triplets (Fiez et al., 1995). Activation of the left IFG was observed in a study by Shahin et al. (2006) that combined functional MRI (fMRI) and ERP and used two target detection paradigms similar to that used in the current study: (1) a semantic task of detecting infrequent word targets denoting animals in a stream of words denoting non-animate objects; (2) a voice gender task detecting infrequent tokens spoken by males in a stream of words spoken by females. Results from fMRI were used to constrain possible anatomical source generators of the ERP. Activation of the IFG in the left hemisphere was seen in the semantic task performed with fMRI, and was associated with negative ERP components to both target and non-target words. Further, responses to targets were larger than responses to non-targets. Peak latencies of these negative ERP components were 450 and 600 ms, respectively, and overlap in time with the high gamma activity observed in the IFG in the present study. These results obtained from neurologically-normal subjects are all concordant with current results, despite the fact that all of our subjects were epileptic patients, and one subject (L275) was trilingual and had non-localizing cognitive deficiencies.

An important distinction between the responses located on PLST and IFG is that activity within pars opercularis of the IFG could vary as a function of behavioral performance (see **Figure 6**). Activity recorded during correctly identified targets was larger than when the target was missed. Further, activity during trials with shorter reaction times peaked earlier than activity during trials when reaction times were longer. This relationship with behavioral performance mirrors that seen in ventrolateral prefrontal cortex of macaques performing a phonemic discrimination task (Russ et al., 2008), and, as discussed above, contrasts with neural activity observed in field AL (Tsunada et al., 2011). The transformation in response characteristics from temporal to frontal lobe is parsimonious with the view that PLST is involved in the process of word object formation, while IFG is involved in the process of word object selection (Shinn-Cunningham, 2008).

MFG appears to also be involved in object selection, as it too responded only to targets (see **Figure 5**) and relevant non-targets during semantic categorization tasks (see Supplementary Figure 5). This activity began later than that in STG and IFG, yet preceded behavioral responses. Activation of the left MFG during a semantic target detection task has been reported using fMRI (Shahin et al., 2006). Variability in responses to targets and relevant non-targets has also been shown in detection tasks using visual stimuli (Kirino et al., 2000; Kiehl et al., 2001; Bledowski et al., 2004; Hampshire et al., 2007; Hon et al., 2012). To varying degrees, MFG as well as IFG were shown to respond either selectively to visual targets or to both targets and relevant non-targets. Additional work will be required to determine the sources of variability that characterized responses during the semantic classification tasks in IFG and MFG.

Strong task-related modulation of high gamma power outside classically defined auditory cortex is consistent with that seen in both the auditory and visual modalities in human ERP and fMRI studies (Sabri et al., 2013; Bansal et al., 2014). In the one study that compared responses to detected vs. undetected sound targets (Sabri et al., 2013), greater activation (as revealed by fMRI) was noted in the parietal lobe, thalamus and basal ganglia. While these regions were not examined in the current study, present results indicate that activity within IFG and MFG (as revealed by high gamma ERBP) is also related to the behavioral outcomes of the task, including the presence of the behavioral response and its timing.

CONCLUDING REMARKS

The response patterns described here reflect multiple processing stages of word object formation that constitute lexical encoding. At a neuroanatomical level, it does not appear that object formation occurs in posteromedial HG. Responses within this region are dominated by representation of the acoustic attributes of speech, and are therefore prelexical. Activity on PLST is also prelexical, but, in contrast to posteromedial HG, can also be strongly modulated by higher-order areas subserving lexical and semantic processing. The modulation on PLST during semantic classification tasks indicates that this region represents an early stage in word object formation.

It should be acknowledged that the subjects that participated in this study are patients who have neurologic deficits, including those in the language domain, and who have been treated with multiple anticonvulsant drugs over long periods of time. This calls into question as to whether findings in this population can be generalized to subjects without neurologic deficits. Despite this limitation, intracranial investigations of neurosurgical patients have been highly fruitful in defining organizational features of auditory and auditory-related cortex (e.g., Crone et al., 2001; Sahin et al., 2009; Chang et al., 2010; Mesgarani and Chang, 2012; Mesgarani et al., 2014). Findings described in the present report confirm and extend our own previous intracranial results demonstrating that PLST exhibits task-related modulation of high gamma activity regardless of behavioral outcome (Nourski et al., 2014a). Finally, results are congruent with non-invasive human studies (e.g., Pugh et al., 1996; Shahin et al., 2006; Gutschalk et al., 2008; Obleser and Eisner, 2009; Leaver and Rauschecker, 2010; Leicht et al., 2010; Simon, 2014) and relevant investigations using experimental animals (e.g., Russ et al.,

2008; Brosch et al., 2011; Tsunada et al., 2011; David et al., 2012; Steinschneider et al., 2013; Sutter et al., 2013).

Future intracranial studies must corroborate current observations and extend them by examining task-related activity in other brain regions known to be important for sound processing. Specifically, investigation of response profiles in anterolateral HG, planum temporale, anterior STG, superior temporal sulcus and MTG will help identify additional stages of word object formation. Similarly, additional work will be needed to further characterize the roles of IFG and MFG in both dominant and non-dominant hemispheres in word object selection. Finally, future studies should include investigation of dynamic interactions between cortical regions, including feedback from higher-order cortices onto sensory areas. This will likely require examination of long-range phase coherence at multiple frequency bands (e.g., theta-gamma) that are likely important in long-range interactions between spatially disparate regions. As we continue investigation of these circuits, our conclusions will undoubtedly be refined and, hopefully, translationally relevant for the understanding of normal speech processing and its disfunction occurring in developmental language disorders, and acquired disorders such as stroke and normal aging.

ACKNOWLEDGMENTS

We thank Haiming Chen, Rachel Gold and Christopher Kovach for help with data collection and analysis. This study was supported by NIH R01-DC04290, UL1RR024979, Hearing Health Foundation and the Hoover Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fnins.2014.00240/abstract

REFERENCES

Alain, C., and Winkler, I. (2012). "Recording event-related brain potentials: application to study auditory perception," in *The Human Auditory Cortex*, eds D. Poeppel, T. Overath, A. N. Popper, and R. R. Fay (New York, NY: Springer Science+Business Media, LLC), 199–224.

Bansal, A. K., Madhavan, R., Agam, Y., Golby, A., Madsen, J. R., and Kreiman, G. (2014). Neural dynamics underlying target detection in the human brain. *J. Neurosci.* 34, 3042–3055. doi: 10.1523/JNEUROSCI.3781-13.2014

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

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300.

Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P. E., Giard, M. H., and Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *J. Neurosci.* 27, 9252–9261. doi: 10.1523/JNEUROSCI.1402-07.2007

Bledowski, C., Prvulovic, D., Hoechstetter, K., Scherg, M., Wibral, M., Goebel, R., et al. (2004). Localizing P300 generators in visual target and distractor processing: a combined event-related potential and functional magnetic resonance imaging study. *J. Neurosci.* 24, 9353–9360. doi: 10.1523/JNEUROSCI.1897-04.2004

Boatman, D. (2004). Cortical bases of speech perception: evidence from functional lesion studies. *Cognition* 92, 47–65. doi: 10.1016/j.cognition.2003.09.010

Brosch, M., Selezneva, E., and Scheich, H. (2011). Representation of reward feedback in primate auditory cortex. *Front. Syst. Neurosci.* 5:5. doi: 10.3389/fnsys.2011.00005

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

Brugge, J. F., Volkov, I. O., Oya, H., Kawasaki, H., Reale, R. A., Fenoy, A., et al. (2008). Functional localization of auditory cortical fields of human: click-train stimulation. *Hear. Res.* 238, 12–24. doi: 10.1016/j.heares.2007.11.012

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Chelazzi, L., Duncan, J., Miller, E. K., and Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol.* 80, 2918–2940.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cereb. Cortex* 11, 761–772. doi: 10.1093/cercor/11.8.761

Crone, N. E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* 112, 565–582. doi: 10.1016/S1388-2457(00)00545-9

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2144–2149. doi: 10.1073/pnas.1117717109

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D. Jr., Redfern, B. B., and Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92, 145–177. doi: 10.1016/j.cognition.2003.11.002

Fiez, J. A., Raichle, M. E., Miezin, F. M., Petersen, S. E., Tallal, P., and Katz, W. F. (1995). PET studies of auditory and phonological processing: effects of stimulus characteristics and task demands. *J. Cogn. Neurosci.* 7, 357–375. doi: 10.1162/jocn.1995.7.3.357

Flinker, A., Chang, E. F., Barbaro, N. M., Berger, M. S., and Knight, R. T. (2010). Sub-centimeter language organization in the human temporal lobe. *Brain Lang.* 117, 103–109. doi: 10.1016/j.bandl.2010.09.009

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what?" Brain-based decoding of human voice and speech. *Science* 322, 970–973. doi: 10.1126/science.1164318

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA: Linguistic Data Consortium.

Griffiths, T. D., Micheyl, C., and Overath, T. (2012). "Auditory object analysis," *in The Human Auditory Cortex*, eds D. Poeppel, T. Overath, A.N. Popper, and R.R. Fay (New York, NY: Springer Science+Business Media, LLC), 199–224.

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nat. Rev. Neurosci.* 5, 887–892. doi: 10.1038/nrn1538

Gutschalk, A., Micheyl, C., and Oxenham, A. J. (2008). Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biol.* 6:e138. doi: 10.1371/journal.pbio.0060138

Hackett, T. A. (2007). "Organization and correspondence of the auditory cortex of humans and nonhuman primates," in *Evolution of Nervous Systems Vol. 4: Primates*, ed J. H. Kaas (New York, NY: Academic Press), 109–119.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hampshire, A., Duncan, J., and Owen, A. M. (2007). Selective tuning of the blood oxygenation level-dependent response during simple target detection dissociates human frontoparietal subregions. *J. Neurosci.* 27, 6219–6223. doi: 10.1523/JNEUROSCI.0851-07.2007

Hickok, G. (2009). The functional neuroanatomy of language. *Phys. Life Rev.* 6, 121–143. doi: 10.1016/j.plrev.2009.06.001

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111. doi: 10.1121/1.411872

Hon, N., Ong, J., Tan, R., and Yang, T. H. (2012). Different types of target probability have different prefrontal consequences. *Neuroimage* 59, 655–662. doi: 10.1016/j.neuroimage.2011.06.093

Hon, N., Thompson, R., Sigala, N., and Duncan, J. (2009). Evidence for long-range feedback in target detection: detection of semantic targets modulates activity in early visual areas. *Neuropsychologia* 47, 1721–1727. doi: 10.1016/j.neuropsychologia.2009.02.011

Howard, M. A. 3rd., Volkov, I. O., Granner, M. A., Damasio, H. M., Ollendieck, M. C., and Bakken, H. E. (1996). A hybrid clinical-research depth electrode for acute and chronic *in vivo* microelectrode recording of human brain neurons. Technical note. *J. Neurosurg.* 84, 129–132. doi: 10.3171/jns.1996.84.1.0129

Howard, M. A., Volkov, I. O., Mirsky, R., Garell, P. C., Noh, M. D., Granner, M., et al. (2000). Auditory cortex on the human posterior superior temporal gyrus. *J. Comp. Neurol.* 416, 79–92. doi: 10.1002/(SICI)1096-9861(20000103)416:1<79::AID-CNE6>3.0.CO;2-2

Indefrey, P., and Cutler, A. (2004). "Prelexical and lexical processing in listening," in *The Cognitive Neurosciences III*, ed M. Gazzaniga (Cambridge, MA: MIT Press), 759–774.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Kiehl, K. A., Laurens, K. R., Duty, T. L., Forster, B. B., and Liddle, P. F. (2001). Neural sources involved in auditory target detection and novelty processing: an event-related fMRI study. *Psychophysiology* 38, 133–142. doi: 10.1111/1469-8986.3810133

Kirino, E., Belger, A., Goldman-Rakic, P., and McCarthy, G. (2000). Prefrontal activation evoked by infrequent target and novel stimuli in a visual target detection task: an event-related functional magnetic resonance imaging study. *J. Neurosci.* 20, 6612–6618.

Lachaux, J. P., Axmacher, N., Mormann, F., Halgren, E., and Crone, N. E. (2012). High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Prog. Neurobiol.* 98, 279–301. doi: 10.1016/j.pneurobio.2012.06.008

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Leicht, G., Kirsch, V., Giegling, I., Karch, S., Hantschk, I., Möller, H. J., et al. (2010). Reduced early auditory evoked gamma-band response in patients with schizophrenia. *Biol. Psychiatry* 67, 224–231. doi: 10.1016/j.biopsych.2009.07.033

Liégeois-Chauvel, C., Musolino, A., Badier, J. M., Marquis, P., and Chauvel, P. (1994). Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. *Electroencephalogr. Clin. Neurophysiol.* 92, 204–214. doi: 10.1016/0168-5597(94)90064-7

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.* 123, 899–909. doi: 10.1121/1.2816572

Nelken, I. (2008). Processing of complex sounds in the auditory system. *Curr. Opin. Neurobiol.* 18, 413–417. doi: 10.1016/j.conb.2008.08.014

Nelken, I., and Bar-Yosef, O. (2008). Neurons and objects: the case of auditory cortex. *Front. Neurosci.* 2, 107–113. doi: 10.3389/neuro.01.009.2008

Niwa, M., Johnson, J. S., O'Connor, K. N., and Sutter, M. L. (2012). Activity related to perceptual judgment and action in primary auditory cortex. *J. Neurosci.* 32, 3193–3210. doi: 10.1523/JNEUROSCI.0767-11.2012

Nourski, K. V., and Brugge, J. F. (2011). Representation of temporal sound features in the human auditory cortex. *Rev. Neurosci.* 22, 187–203. doi: 10.1515/rns.2011.016

Nourski, K. V., Brugge, J. F., Reale, R. A., Kovach, C. K., Oya, H., Kawasaki, H., et al. (2013). Coding of repetitive transients by auditory cortex on posterolateral superior temporal gyrus in humans: an intracranial electrophysiology study. *J. Neurophysiol.* 109, 1283–1295. doi: 10.1152/jn.00718.2012

Nourski, K. V., and Howard, M. A. 3rd. (2014). Invasive recordings in human auditory cortex. *Hand. Clin. Neurol.* (in press).

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

Nourski, K. V., Steinschneider, M., Oya, H., Kawasaki, H., and Howard, M. A. 3rd. (2014a). Modulation of response patterns in human auditory cortex during a target detection task: an intracranial electrophysiology study. *Int. J. Psychophysiol.* doi: 10.1016/j.ijpsycho.2014.03.006. [Epub ahead of print].

Nourski, K. V., Steinschneider, M., Oya, H., Kawasaki, H., Jones, R. D., and Howard, M. A. 3rd. (2014b). Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. *Cereb. Cortex* 24, 340–352. doi: 10.1093/cercor/bhs314

Obleser, J., and Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends Cogn. Sci.* 13, 14–19. doi: 10.1016/j.tics.2008.09.005

Petkov, C. I., and Jarvis, E. D. (2012). Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front. Evol. Neurosci.* 4:12. doi: 10.3389/fnevo.2012.00012

Petkov, C. I., Kang, X., Alho, K., Bertrand, O., Yund, E. W., and Woods, D. L. (2004). Attentional modulation of human auditory cortex. *Nat. Neurosci.* 7, 658–663. doi: 10.1038/nn1256

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

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Fulbright, R. K., Byrd, D., Skudlarski, P., et al. (1996). Auditory selective attention: an fMRI investigation. *Neuroimage* 4, 159–173. doi: 10.1006/nimg.1996.0067

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

Reddy, C. G., Dahdaleh, N. S., Albert, G., Chen, F., Hansen, D., Nourski, K., et al. (2010). A method for placing Heschl gyrus depth electrodes. *J. Neurosurg.* 112, 1301–1307. doi: 10.3171/2009.7.JNS09404

Russ, B. E., Orr, L. E., and Cohen, Y. E. (2008). Prefrontal neurons predict choices during an auditory same-different task. *Curr. Biol.* 18, 1483–1488. doi: 10.1016/j.cub.2008.08.054

Sabri, M., Humphries, C., Verber, M., Mangalathu, J., Desai, A., Binder, J. R., et al. (2013). Perceptual demand modulates activation of human auditory cortex in response to task-irrelevant sounds. *J. Cogn. Neurosci.* 25, 1553–1562. doi: 10.1162/jocn_a_00416

Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within Broca's area. *Science* 326, 445–449. doi: 10.1126/science.1174481

Shahin, A. J., Alain, C., and Picton, T. W. (2006). Scalp topography and intracerebral sources for ERPs recorded during auditory target detection. *Brain Topogr.* 19, 89–105. doi: 10.1007/s10548-006-0015-9

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Simon, J. Z. (2014). The encoding of ausitory objects in auditory cortex: insights from magnetoencephalography. *Int. J. Psychophysiol.* doi: 10.1016/j.ijpsycho.2014.05.005. [Epub ahead of print].

Steinschneider, M., Fishman, Y. I., and Arezzo, J. C. (2008). Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cereb. Cortex* 18, 610–625. doi: 10.1093/cercor/bhm094

Steinschneider, M., Nourski, K. V., and Fishman, Y. I. (2013). Representation of speech in human auditory cortex: is it special? *Hear. Res.* 305, 57–73. doi: 10.1016/j.heares.2013.05.013

Steinschneider, M., Nourski, K. V., Kawasaki, H., Oya, H., Brugge, J. F., and Howard, M. A. 3rd. (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb. Cortex* 21, 2332–2347. doi: 10.1093/cercor/bhr014

Steinschneider, M., Schroeder, C. E., Arezzo, J. C., and Vaughan, H. G. Jr. (1994). Speech-evoked activity in primary auditory cortex: effects of voice onset time. *Electroencephalogr. Clin. Neurophysiol.* 92, 30–43. doi: 10.1016/0168-5597(94)90005-1

Sutter, M., O'Connor, K. N., Downer, J., Johnson, J., and Niwa, M. (2013). Hierarchical effects of attention on amplitude modulation encoding in auditory cortex. *J. Acoust. Soc. Am.* 134, 4085. doi: 10.1121/1.4830920

Travis, K. E., Leonard, M. K., Chan, A. M., Torres, C., Sizemore, M. L., Qu, Z., et al. (2013). Independence of early speech processing from word meaning. *Cereb. Cortex* 23, 2370–2379. doi: 10.1093/cercor/bhs228

Tsunada, J., Lee, J. H., and Cohen, Y. E. (2011). Representation of speech categories in the primate auditory cortex. *J. Neurophysiol.* 105, 2634–2646. doi: 10.1152/jn.00037.2011

Winkler, I., van Zuijen, T. L., Sussman, E., Horváth, J., and Näätänen, R. (2006). Object representation in the human auditory system. *Eur. J. Neurosci.* 24, 625–634. doi: 10.1111/j.1460-9568.2006.04925.x

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

Received: 29 May 2014; accepted: 22 July 2014; published online: 11 August 2014.
Citation: Steinschneider M, Nourski KV, Rhone AE, Kawasaki H, Oya H and Howard MA III (2014) Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Front. Neurosci. ***8****:240. doi: 10.3389/fnins.2014.00240*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 30 September 2014
doi: 10.3389/fnins.2014.00306

Sensitivity of human auditory cortex to rapid frequency modulation revealed by multivariate representational similarity analysis

Marc F. Joanisse* and Diedre D. DeSouza

Department of Psychology, Brain and Mind Institute, The University of Western Ontario, London, ON, Canada

Edited by:
Einat Liebenthal, Brigham and Women's Hospital, USA

Reviewed by:
James W. Lewis, West Virginia University, USA
Iiro P. Jääskeläinen, Aalto University, Finland
Thomas Talavage, Purdue University, USA

****Correspondence:***
Marc F. Joanisse, Department of Psychology, Brain and Mind Institute, The University of Western Ontario, London, Ontario, ON N6K 4B1, Canada
e-mail: marcj@uwo.ca

Functional Magnetic Resonance Imaging (fMRI) was used to investigate the extent, magnitude, and pattern of brain activity in response to rapid frequency-modulated sounds. We examined this by manipulating the direction (rise vs. fall) and the rate (fast vs. slow) of the apparent pitch of iterated rippled noise (IRN) bursts. Acoustic parameters were selected to capture features used in phoneme contrasts, however the stimuli themselves were not perceived as speech *per se*. Participants were scanned as they passively listened to sounds in an event-related paradigm. Univariate analyses revealed a greater level and extent of activation in bilateral auditory cortex in response to frequency-modulated sweeps compared to steady-state sounds. This effect was stronger in the left hemisphere. However, no regions showed selectivity for either rate or direction of frequency modulation. In contrast, multivoxel pattern analysis (MVPA) revealed feature-specific encoding for direction of modulation in auditory cortex bilaterally. Moreover, this effect was strongest when analyses were restricted to anatomical regions lying outside Heschl's gyrus. We found no support for feature-specific encoding of frequency modulation rate. Differential findings of modulation rate and direction of modulation are discussed with respect to their relevance to phonetic discrimination.

Keywords: frequency modulation, auditory cortex, heschl's gyrus, multivoxel pattern analysis (MVPA), functional magnetic resonance imaging (fMRI)

INTRODUCTION

During verbal communication, our auditory system is charged with the task of sorting through a complex acoustic stream in order to identify relevant stimulus features, and then integrating this information into a unified phonetic percept that can allow us to perceive the incoming message. This process occurs amidst competing sources of information and significant variability in how a given speech sound is produced both within- and between-speakers. Yet humans can decode auditory speech both accurately and in a way that usually seems effortless.

A key characteristic of the speech signal is that it contains acoustic complexities in both the spectral and temporal domains. Spectrally, it contains simultaneous bands of high and low intensities across a range of frequencies. Temporally, the signal is amplitude modulated such that its intensity is rapidly changing and fast fading, and it is frequency modulated so that spectral information changes at a rapid rate. This multicomponent nature of the acoustic speech signal makes it unique in the domain of auditory processing.

In the present study we focus on the neural processing of one specific characteristic of temporal-acoustic speech processing, namely rapid frequency modulation (FM). The production of many phonemes results in a concentration of resonant frequencies, known as formants. The frequencies of these formants will vary depending on the configuration of the vocal tract during articulation. Because speech is produced in a dynamic fashion, formant frequencies tend to rapidly change at differing rates over time (Hillenbrand et al., 1995). Accordingly, manipulating the FM characteristics of formants in speech changes its perceived phonemic characteristics (e.g., Stevens and Klatt, 1974). For example, slowing the rate of a syllable-initial stop consonant's formant transitions will change the perception of /b/ to /w/ (Liberman et al., 1956). Likewise, changing the direction of the second formant's (F2) transition will change a /b/ to /d/ (Miller and Liberman, 1979). Given the important role of formant transitions in speech perception, the present research focuses on examining the neural underpinnings of how these FM acoustic cues are perceived.

Prior work supports the view that auditory cortex in superior temporal gyrus (STG) is organized in a hierarchical fashion that supports the processing of increasingly complex characteristics of auditory signals. Thus, as we move outward from the core region of auditory cortex formed by Heschl's gyrus toward the "belt" and "parabelt" regions that surround it, we observe regions that respond to the increasingly complex spectral and temporal characteristics of acoustic stimuli. Converging support for this notion has come from studies of auditory cortex in humans (Wessinger et al., 2001; Chevillet et al., 2011) and non-humans (Rauschecker et al., 1995; Kaas et al., 1999; Kikuchi et al., 2010), and across a variety of imaging modalities including functional

magnetic resonance imaging (fMRI), magnetoencephalography (MEG) and electrophysiology (Mendelson et al., 1993; Tian and Rauschecker, 2004; Godey et al., 2005; Heinemann et al., 2010; Carrasco and Lomber, 2011).

Studies of the sort do seem to have some implications for how the acoustic form of speech is processed. For instance, Chevillet et al. (2011) compared neural responses to sounds of increasing spectral complexity, namely pure tones, broadband noise, and vowel sounds. They found that pure tones elicited activation in Heschl's gyrus, whereas broadband noise elicited activation in both auditory core as well as belt areas both medial and lateral to the auditory core. Lastly, vowel sounds elicited activation in core, belt, and parabelt regions that surround them. This indicates both a greater sensitivity to spectrally complex sounds in primary auditory cortex, and the increasing recruitment of surrounding brain areas as this complexity increases. Note that although the literature generally supports the notion of a hierarchy from core to belt in auditory cortex, there is some suggestion that primary auditory cortex does itself contain regions sensitive to higher-order auditory scenes (Nelken et al., 2003). Thus, one cannot discount the possibility that this region can decode auditory events as complex objects for subsequent recognition.

FREQUENCY MODULATION IN HUMAN AUDITORY CORTEX

Most of the studies described above have focused on the effect of modulating the spectral complexity of sounds in order to describe the function of primary vs. secondary auditory cortex. Consequently, much less is known about the organization of auditory cortex with respect to rapid temporal FM cues that are also important for speech. Most of what we know about the coding of FM features in auditory cortex comes from single- and multi-unit electrode recordings of auditory cortex in non-humans. These studies have identified evidence of neuronal selectivity to FM vs. acoustically similar steady-state sounds, across several animal species (Mendelson et al., 1993; Nelken and Versnel, 2000; Liang et al., 2002; Washington and Kanwal, 2008; Kusmierek and Rauschecker, 2009). Moreover, neurons may be individually tuned to specific characteristics of these FM sounds. For instance, Mendelson et al. identified neurons in the primary auditory cortex of cats that are systematically distributed according to either the rate and direction of frequency modulation sweeps. Such findings raise the possibility that auditory cortex in humans is also organized in a way that is preferentially sensitive to these aspects of FM sounds.

Neuroimaging studies in humans have also identified regions of auditory cortex that show a preference to time-varying sounds. For instance, Zatorre and Belin (2001) used positron emission tomography (PET) to examine both the spectral and temporal variation of sounds within human auditory cortex by playing sequences of steady-state pure tones of differing frequencies or durations. The authors found bilateral activation of the core auditory cortex as the rate of pitch variation increased, and bilateral activation of anterior STG in response to spectral variation. Additionally, they found that activation in response to the temporal manipulation was left lateralized while responses to the spectral manipulation were right lateralized. Similarly, Hall and colleagues (Hall et al., 2002) found enhanced fMRI response in STG for FM tone complexes compared to acoustically similar static tones. These FM-sensitive regions included Heschl's gyrus in the left hemisphere, and STG regions adjacent to Heschl's gyrus bilaterally.

There is also some reason to believe that auditory cortex sensitivity to frequency modulation is related to speech processing. Joanisse and Gati (2003) used fMRI to examine activation in superior temporal cortex in response to sequences of stop consonants that varied in their rapid FM characteristics, or vowels that varied in terms of steady-state spectral characteristics. A pair of control conditions used sets of pure tones that also differed along either FM or spectral dimensions. They found that consonants and FM tones yielded stronger activation in left STG and surrounding areas, whereas a congruent effect in the right hemisphere was observed for vowel and pure tone pitch discrimination. The findings again suggest some special status for FM processing in auditory cortex, and that this effect is generally left lateralized.

That said, such studies leave open the possibility that humans maintain cortical regions within primary or secondary auditory cortex that are specially tuned to individual FM features of sounds. Recently Hsieh et al. (Hsieh et al., 2012) examined this issue in humans using fMRI. Their study presented tone complexes that varied in the rate and direction of frequency change; stimuli involved either shorter or longer complex tone sweeps that were either rising or falling. Interestingly, they did not identify brain regions that robustly differentiated either of these two dimensions, suggesting that auditory cortex is not topographically organized in a way that differentiates either the rate nor the direction of FM; that is, no region was more sensitive to rising than falling tones, or showed enhanced activation for faster vs. slower rates of modulation. However, the results were different when the authors employed a multivoxel pattern analysis (MVPA) approach, which takes into account the overall pattern of voxel activity for each stimulus type. This analysis identified (in a subset of subjects) unique patterns of activation for both the rate and direction of FM sweeps in primary auditory cortex and surrounding regions of STG. This suggests that FM-selective brain regions do exist in humans, but that they occur on a level of grain that is much smaller than what can be identified using typical univariate fMRI approaches.

That said, it is not clear how this result bears more narrowly on the question of FM cues for phonemic processing. The stimuli in the Hsieh et al. study involved contrasting relatively slow-going rate changes (0.83 and 3.3 octaves per second) that are not on the order of those used in formant transitions that cue phonemic speech contrasts. Instead the contrasts are more similar to those that signal lexical tone contrasts in some languages; indeed, listeners in their study were native speakers of Mandarin, a tonal language. In addition, the way in which the rate variable was manipulated in this earlier study merits some discussion. FM rate can be modified in three possible ways: the rate of change over time, the extent of frequency change, and the duration of the stimulus itself. However, it is not possible to manipulate one of these independently, and thus two of three factors will always be confounded. Thus, in the Hsieh et al. study the length of the stimulus was manipulated to yield fast vs. slow FM sweeps, such

that rate was confounded with overall stimulus duration. This raises the possibility that differences in neural responses to rate reflected sensitivity to the duration of the stimulus rather than the rate of modulation itself. To be clear, such confounds are likely a necessary element of FM stimuli, however it does leave open the possibility that different results could occur when the stimulus rate parameter is manipulated differently.

MOTIVATION FOR THE PRESENT STUDY

Our central focus in the present study was to examine the neural processing of rapid FM features in non-speech acoustic stimuli, compared to acoustically similar steady-state sounds. The intention was to examine how the human brain processes and differentiates characteristics of these stimuli and, in particular, whether different subregions of auditory cortex respond preferentially to these specific features. Consistent with prior studies, FM stimuli in general should yield greater activation both in primary auditory cortex and surrounding regions, when compared to steady-state sounds of similar spectral complexity (Rauschecker et al., 1995; Kusmierek and Rauschecker, 2009). The effect should also be stronger in the left hemisphere. Additionally, we adopted a design that examined differences in neural response to specific features of FM, specifically the direction and rate of change in frequency. This allowed us to assess whether subregions of auditory cortex are specifically tuned to basic features of rapid-FM sounds.

Central to our approach was using stimuli that capture key acoustic features of speech. Thus, the two FM modulation rates we used roughly correspond to the duration of second-formant transitions observed in stop consonants and semivowels (e.g., /ba/ vs. /wa/; see **Figure 1B**). We also sought to capture the spectrotemporal complexity of speech by employing iterative rippled noise (IRN) stimuli. IRN is a type of broadband noise that maintains the types of discernible spectral and temporal regularities that are usually associated with narrowband tones (Swaminathan et al., 2008). Just as importantly, IRN does not contain phonetic cues, and does not yield speech-like auditory illusions. This allowed us to capture the general spectral complexity typical of speech, while preserving the ability to manipulate perceived pitch and therefore the FM characteristics of stimuli. IRN stimuli were useful here because they are both spectrally broadband and can contain temporal features mimicking phoneme contrasts, but they are not perceived as speech *per se*. Their spectral characteristics are especially relevant to this end; past research has demonstrated that regions within auditory cortex respond differentially to speech vs. spectrally simple non-speech sounds such as tones (Binder et al., 2000; Whalen et al., 2006). Likewise, IRN stimuli can simulate the high harmonics-to-noise ratio (HNR) of speech (Boersma, 1993). HNR is a higher-order acoustic attribute that indexes the harmonic structure of sounds, and which tends to be higher in natural vocalizations than other types of environmental sounds. There is evidence that subregions of core and belt auditory cortex are specifically tuned to this characteristic due to the increased recruitment of neurons sensitive to multiple frequency-combinations (Lewis et al., 2005). Likewise, there appears to be strong overlap in auditory cortical activity in response to IRN and human vocalizations that is directly attributable to their similarity in HNR (Lewis et al., 2009).

We chose to study FM processing using IRN sounds rather than actual phonetic stimuli in order to avoid potential extraneous influences of speech on the resulting fMRI activation patterns. That is, intelligible speech both comprises acoustic-phonetic information and conveys meaning. Thus, it is challenging to differentiate neural responses to the acoustic features of speech from the effects of its articulatory-phonetic and semantic content.

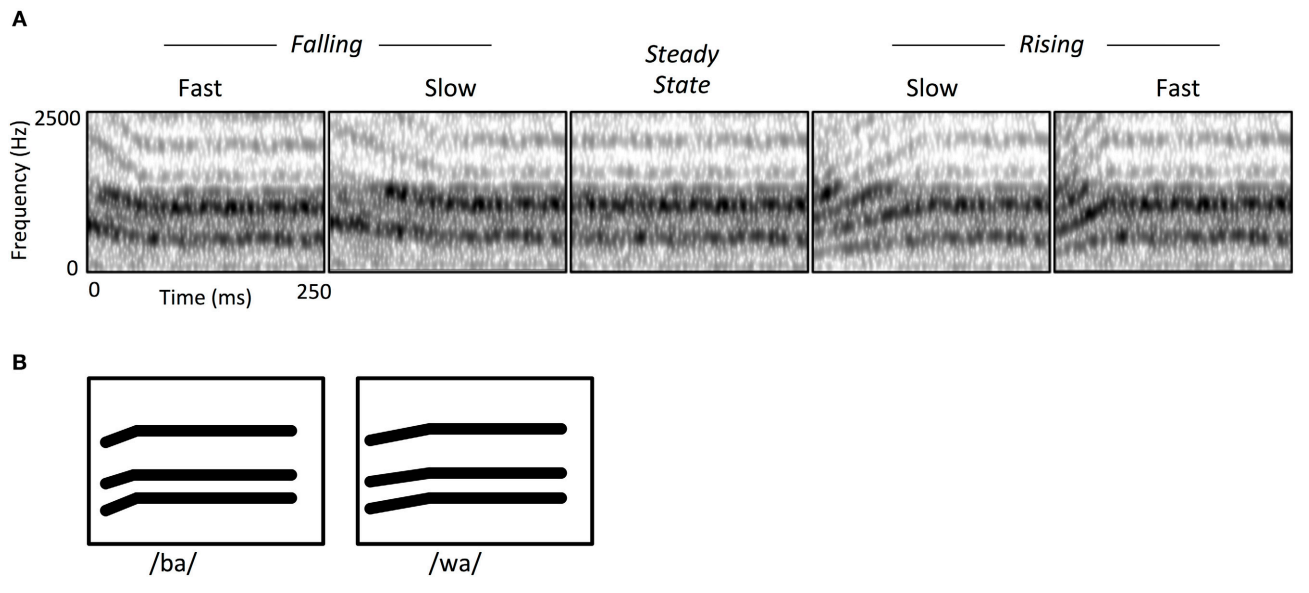


FIGURE 1 | (A) Spectrograms of iterative rippled noise (IRN) stimuli, showing frequency modulated (FM) sweeps and steady state conditions. FM stimuli consisted of rising or falling sweeps, at either a fast or short rate of change. All FM stimuli were followed by a steady state portion to hold all stimulus durations at 250 ms. **(B)** Schematized spectrogram of a speech contrast (/ba/ and /wa/) illustrating how the IRN stimuli simulate transition rate contrasts of formant contours in natural speech sounds. Note that while this phoneme pair illustrates a rising contrast, other speech sounds can instead involve falling contrasts.

Speech that is intelligible evokes activation in a broader portion of temporal cortex than speech stimuli that have been distorted to the point of unintelligibility (Scott et al., 2000). Likewise, when sinewave tones are systematically combined to approximate the center frequencies of speech formants (i.e., sinewave speech; Remez et al., 1981), listeners can perceive them as having phonetic content. Accordingly, different patterns of activation are observed in temporal cortex when listeners perceive these sinewave sounds as phonetic, compared to when they do not (Liebenthal et al., 2003; Möttönen et al., 2006). Overall, using IRN stimuli allowed us to isolate neural effects of FM processing from effects that occur in response to semantic integration or articulatory-phonetic processing.

We employed two statistical approaches to examine FM processing in auditory cortex. In addition to standard univariate analyses we also used a multivariate approach of representational similarity analysis (RSA). This is an MVPA methodology that computes the similarity of voxel activation patterns among different experimental conditions. While conventional univariate neuroimaging analyses are useful for detecting regional activation differences, they do not provide any information regarding representational differences that occur at a grain of analysis below that afforded by fMRI voxel sizes. On the other hand, MVPA approaches allow us to detect activation patterns in regions of interest even when average activation is similar across conditions (Kriegeskorte et al., 2008). It was therefore expected that RSA could reveal representational differences among FM features in auditory cortex even if univariate analyses failed to reveal large-scale differences in the degree or extent of fMRI activation.

METHODS

SUBJECTS

Sixteen neurologically healthy adult participants were recruited for this study (eight female, eight male); mean age was 27 years (range 18–31 years). All participants were right-handed, monolingual native English speakers with normal audition by self-report. Informed consent was obtained from each participant in accordance with the University of Western Ontario Medical Research Ethics Board.

STIMULI

The auditory stimuli consisted of Iterative Rippled Noise (IRN) bursts, which are broadband noise manipulated in a way that produces a perceived pitch contour while maintaining wideband spectral complexity (**Figure 1A**). Stimuli were created in Matlab (MathWorks, 2010) at a 44.1 KHz sample rate (16-bit quantization), matching the procedure from Swaminathan et al. (2008), whereby a noise impulse is delayed and added to the sample at each iteration, with a delay of 4 ms and a gain of 1. For each stimulus we created a pitch contour represented by a polynomial equation and then created a time varying IRN stimulus that mimicked that input pitch contour by modulating the time delay at each iteration. There were four FM stimulus sweeps in which the center frequency of the IRN was varied linearly over time: Rise-Fast, Rise-Slow, Fall-Fast, and Fall-Slow (**Table 1**, **Figure 1A**). The "Fast" sweep had an FM rate of 20 octaves/s and a duration of 50 ms; the FM rate in the "Slow" condition was 10 octaves/s and a 100 ms duration. Note that our goal was to maintain the same duration for all stimuli, which should at least partially overcome the concern that different sweep rates necessarily require either different durations or frequency extents for a stimulus. For that reason, an additional steady-state period was added to the end of each sweep, yielding a total stimulus duration of 250 ms (**Figure 1A**, **Table 1**). We also created a fifth "Steady-State" stimulus condition which consisted of an IRN of the same duration and intensity as the FM stimuli, but which had a constant perceived frequency of 1200 Hz.

Table 1 | Acoustic characteristics of the IRN stimuli, showing center frequency contours (Hz) for the frequency modulated (FM) and steady-state stimuli.

	Time (*ms*)			
Condition	**0**	**50**	**100**	**250**
Rise-Fast	600	1200	1200	1200
Fall-Fast	1800	1200	1200	1200
Rise-Slow	600	900	1200	1200
Fall-Slow	1800	1500	1200	1200
Steady-state	1200	1200	1200	1200

During scanning, stimuli were presented binaurally via MR compatible headphones (Sensimetrics Model S14). Participants were instructed to passively listen to the audio stimuli. A silent movie was displayed via a projector to keep the participant alert. We employed an event-related design in which stimuli were presented at randomly jittered SOAs of 2.1, 4.2, 6.3, or 8.4 s (corresponding to integer multiples of the 2.1 s scan repetition time). A sparse scanning paradigm was used in which silent gaps were introduced between each EPI scan, with auditory stimuli presented during these silent gaps, 50 ms following the end of the previous scan to eliminate possible acoustic masking. The scanning session was divided into six runs with brief rests between each. Within each run, stimuli were presented in pseudo-random order, with 19 presentations of each condition, for a total of 95 presentations per condition over the entire session.

NEUROIMAGING

Images were acquired using a 3.0 Tesla Siemens TIM Trio Scanner equipped with a 32-channel head coil. Functional images were acquired in an axial orientation using an iPAT parallel acquisition sequence (GRAPPA, generalized auto-calibrating partially parallel acquisition; acceleration factor 2). Six runs of 252 T2*-weighted functional scans were acquired for each subject (voxel size $= 3 \times 3 \times 3$ mm; FOV $= 192 \times 192$ mm; $TA = 1.6$ s, plus 0.5 s inter-scan gap, yielding an effective $TR = 2.1$ s; $TE = 30$ ms; matrix size: $64 \times 64 \times 28$). Twenty-eight slices per volume were obtained with no inter-slice gap, providing full coverage of temporal and occipital lobes, but only partial coverage of the upper portion of the cerebrum. Specifically, coverage excluded superior portions of the somatosensory cortex, motor cortex and superior parietal lobe. A whole-brain high-resolution T1-weighted anatomical image was also obtained within-session prior to the first functional run using a 3D gradient-echo parallel acquisition

sequence (MPRAGE; GRAPPA acceleration factor = 2; voxel size = 1 × 1 × 1 mm; *TR* = 2.3 s; *TE* = 2.98 ms; Flip angle = 9°; matrix size: 192 × 256 × 256).

UNIVARIATE STATISTICAL ANALYSES

Imaging data were analyzed using the AFNI software package (Cox, 1996). All functional scans were motion corrected using a 3D rigid body transform (AFNI *3dvolreg*) registered to the first functional volume of the first run. Statistical parametric maps were created using a general linear model (GLM, AFNI 3*dDeconvolve*) composed of six regressors; five condition regressors (Fall-Fast, Fall-Slow, Rise-Fast, Rise-Slow, Steady-State), and a single motion parameter estimate calculated as the root mean square of the six movement estimates derived from the motion correction step. Each task predictor was convolved with a canonical hemodynamic response function. Group statistical maps were created by registering each subject-wise map to a standard template (the TT_N27 "Colin" brain template) in the stereotaxic space of Talairach and Tournoux (1988), using an automatic registration procedure (the AFNI *@auto_tlrc* script, least-squares cost function). Each statistical map was then resampled to a resolution of 1 mm^3 and spatially filtered with a 5 mm FWHM Gaussian kernel.

Statistical contrasts were performed via *t*-tests at the group level as follows: the Steady-State condition was contrasted with the four combined FM conditions to identify regions of greater sensitivity to time varying vs. static components of acoustic signals. The second and third contrasts identified voxels sensitive to either the rate or direction of FM sweeps (with Steady State condition set as a condition of no interest). For modulation rate, we contrasted (Rise-Fast + Fall-Fast) vs. (Rise-Slow + Fall-Slow); for sweep direction, we contrasted (Rise-Fast + Rise-Slow) vs. (Fall-Fast + Fall-Slow). Contrasts were thresholded at $p < 0.05$ corrected for multiple comparisons based on a voxel-wise threshold of $p < 0.002$ and a cluster size threshold of 971 mm^3 (estimated using a 10,000-iteration Monte Carlo procedure, accounting for observed mean spatial blurring in each dimension; AFNI *3dClustSim*).

MULTIVARIATE STATISTICAL ANALYSIS

Data were also analyzed using representational similarity analysis (RSA; Kriegeskorte et al., 2008), to examine the relative similarity of the voxel activation pattern across conditions. Analyses were performed within two regions of interest (ROIs): auditory cortex defined functionally across the temporal lobe, and Heschl's gyrus, with both ROIs based on regions defined within a previous study (Linke et al., 2011; see **Figure 2**). The Auditory cortex ROI was defined as regions of temporal cortex that was activated in the Linke et al. study during the encoding, maintenance and comparison of tone stimuli. This ROI subtended the anterior and posterior plane of STG and STS. The Heschl's gyrus ROI was identified anatomically using a standard atlas. Note that the two ROIs were non-overlapping such that the Auditory ROI excluded voxels falling within the Heschl's ROI and vice-versa. We performed RSA separately for activation patterns within the ROIs of the left, right and combined hemispheres, for a total of six analyses.

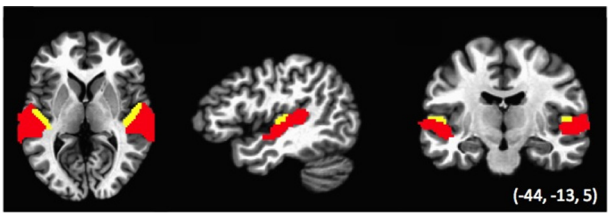


FIGURE 2 | ROIs used in the Relative similarity analysis (RSA). Areas in red (Auditory cortex ROI) are auditory-functionally defined regions of left and right temporal lobe; areas in yellow (Heschl's gyrus ROI) correspond to the anatomically defined Heschl's gyrus in a standard atlas (see Linke et al., 2011). Note that the two sets of ROIs do not overlap.

Voxel activation patterns were computed for each subject on each of the five stimulus types using a GLM as specified above, but with no spatial smoothing, and with separate GLM maps obtained for even and odd runs. This yielded two sets of five statistical maps per subject. Beta coefficients for each statistical map were ROI masked and subjected to Spearman correlations between even and odd runs for each combination of conditions, yielding a 5 × 5 similarity matrix for each subject. Next, statistical contrasts were performed groupwise to investigate the dissimilarity of the two dynamic features of interest. RSA for direction of modulation was assessed by performing a pairwise *t*-test for coefficients in the rising vs. falling conditions, collapsing across the two rate conditions; RSA for rate was assessed by performing a pairwise *t*-test for the fast and slow conditions, collapsing across the two direction directions. Significant differences in an ROI indicated this region differentially encodes information regarding the categories of stimuli under investigation.

RESULTS

UNIVARIATE ANALYSIS

The first contrast investigated the existence of specialized regions for processing FM sounds compared to spectrally similar steady-state sounds, and was computed using a one-sample *t*-test for the Steady-State predictor vs. zero (**Figure 3A**, **Table 2**). Results revealed clusters of activation in bilateral posterior STG in and around Heschl's gyrus. We next contrasted the combined FM conditions vs. the Steady State condition. As indicated in **Figure 3B** and the lower portion of **Table 2**, we found clusters of activation throughout bilateral auditory cortex peaking in portions of superior temporal gyrus (STG) both anterior and posterior to Heschl's gyrus, and extending more ventrally toward superior temporal sulcus (STS). This effect was more pronounced in the left than right hemisphere, taking into account the total size of the two separate clusters in L-STG/STS. A significant cluster was also observed in the right supramarginal gyrus (SMG).

We also sought to identify regions responsible for processing either the direction or rate of FM sweeps. For the effect of rate (fast vs. slow), the two levels of direction were conflated: (Rise-Fast + Fall-Fast) vs. (Rise-Slow + Fall-Slow). In a similar fashion, the effect of direction was examined by collapsing over rates: (Rise-Fast + Rise-Slow) vs. (Fall-Fast + Fall-Slow). Neither of these contrasts yielded significant difference in either direction

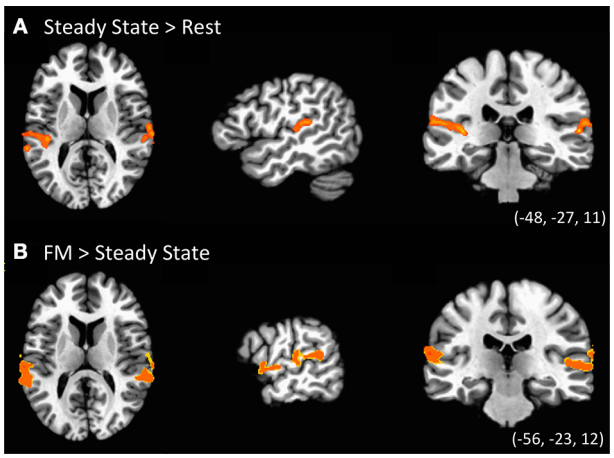


FIGURE 3 | Univariate statistical contrasts. (A) Steady state condition vs. rest reveals bilateral activation in superior temporal gyrus within and around primary auditory cortex. **(B)** Frequency-modulated (FM) sweeps vs. steady state, showing greater activation for FM sounds in the superior temporal plane of both right and left hemispheres. Group statistical maps are superimposed on a standardized template. Voxel-wise threshold: $p < 0.002$, corrected to $p < 0.05$ for a 971 mm^3 cluster extent.

Table 2 | Location and size of the peak voxel activation for the univariate analysis.

Region	Talairach Coordinates			Size (mm^3)
	X	*Y*	*Z*	
STEADY vs. REST				
L STG	−43	34	14	3667
R STG	56	28	8	1404
DYNAMIC > STEADY				
L STG	−61	24	13	2166
L STG	−57	8	3	1514
R STG	56	8	3	3328
R SMG	44	58	34	1360

Corrected $\alpha = 0.05$; voxel-wise threshold: $p < 0.002$; cluster size threshold = 971 mm^3.

at a threshold of significance corrected for multiple comparisons. This lack of effect persisted even when not controlling for cluster extent at this same voxel-wise significance threshold.

MULTIVARIATE ANALYSIS

The RSA analysis measured the similarity of voxel activation patterns for FM rate and direction within a given ROI. RSA matrices and contrasts are visualized in **Figure 4**. Within each matrix in **Figure 4**, the correlations between each grouping of conditions is plotted, with the relative intensity of each square denoting the degree of similarity; statistical analyses then contrasted the correlation coefficients in order to assess whether representational similarity within each ROI was different for the conditions of interest. The first contrast examined whether different directions of FM (rising vs. falling) yielded different patterns of activation. The results revealed strong evidence of direction-specific activity patterns in left and right Auditory ROIs, and in the left Heschl's gyrus ROI. This is best visualized by stronger correlations within each sub-plot for FM conditions along the diagonal (rising vs. rising and falling vs. falling) compared to the off-diagonal (rising vs. falling). In contrast, the RSA analysis did not reveal strong evidence for differentiation within the rate manipulation, marked by a failure to find significantly greater similarity of activation patterns within-category vs. between-category. These results suggest that auditory cortex is generally more sensitive to changes in direction than to changes in the rate of frequency-modulated sweeps for the rapid FM acoustic features explored in this experiment.

DISCUSSION

Spoken language comprises a dynamic and broadband acoustic signal made up of many types of temporal and spectral features. In the present study we were interested in one specific aspect of speech, the rapid temporal frequency modulations that are used to signal phonetic contrasts such as place of articulation. Our stimuli involved non-speech sounds that isolated specific characteristics of frequency modulation (FM) namely direction and rate of frequency changes.

We first investigated whether FM sweeps and steady-state sounds elicited different responses in large-scale brain activity patterns. By contrasting brain regions that were activated in response to the two classes of stimuli (FM sweeps vs. steady-state sounds), we were able to demonstrate that there are indeed differences in both the extent and magnitude of activation within both core and belt auditory cortex. This finding is consistent with prior studies showing that auditory cortex is generally organized in a way that codes for increasing complexity of auditory information as it extends outward from primary auditory cortex to regions that surround it (e.g., Rauschecker et al., 1995; Hall et al., 2002; Chevillet et al., 2011). The present study fits well with such findings, illustrating that this effect can be driven by rapid FM characteristics of sounds. The steady-state sounds used in the current study were as spectrally complex as the FM sweeps; the only difference was the time varying nature of the FM sounds.

We also observed an interesting pattern of lateralization of activation in response to frequency-modulated sweeps, such that the left hemisphere displayed a greater extent of activation than the right hemisphere. This finding lends further support to the role of FM in language processing given the theory that the left hemisphere is specialized for processing the salient auditory features that are the components of more complex acoustic signals such as speech. The finding is in line with previous research that demonstrated that temporal modulation of auditory inputs yield stronger left hemisphere activation compared to steady-state stimuli (e.g., Zatorre and Belin, 2001; Hall et al., 2002) and that congruent effects occur for speech stimuli incorporating these rapid temporal characteristics (Joanisse and Gati, 2003). This suggests that sensitivity to rapid temporal cues reflects a fundamental specialization of left auditory cortex for processing time varying acoustic signals, both for speech and non-speech.

One note about this interpretation is in order however. We also observed a somewhat greater extent of left-hemisphere activation

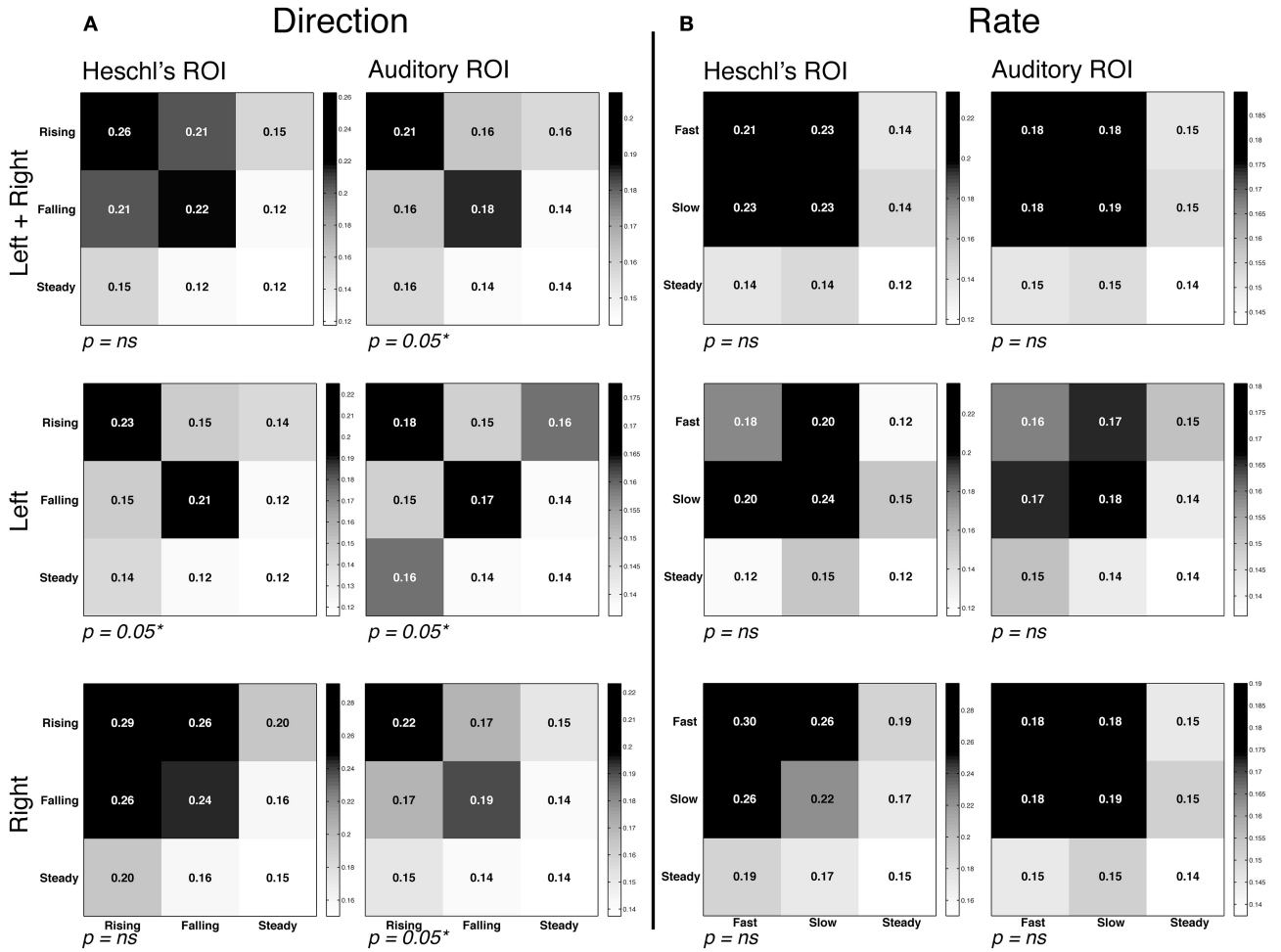


FIGURE 4 | Representational similarity analysis. Mean correlation of activity patterns between even runs (rows) and odd runs (columns), at each region of interest (see **Figure 2**). Within-category correlations correspond to the diagonal of the matrix; between-category correlations correspond to off-diagonal values. Higher similarity is depicted by higher correlations and darker shading. Note that overall correlation ranges varied by ROI, and therefore different intensity scales are used in each ROI in order to highlight differences in the relative degree of similarity. Also note that statistical analyses compared the correlation coefficients of cells, and not the significance of each correlation on its own. Thus, *P*-values correspond to the comparison of between vs. within-condition correlation coefficients for the rapid temporal conditions (the upper left 2 × 2 square of each matrix). In all graphs the third row/column corresponds to the steady state condition. **(A)** Correlation matrices for the direction contrast, collapsing over the rate manipulation. We observed significantly greater representational similarity within the FM conditions (i.e., rising-rising and falling-falling) compared to between conditions (rising-falling/falling-rising). **(B)** Correlation matrices for the rate contrast, collapsing over the direction manipulation. No effect of rate was observed, marked by similar degrees of similarity within condition (fast-fast/slow-slow) and between condition (fast-slow/slow-fast).

for the steady-state condition alone even though no FM cues were present in that case (**Figure 3A**). The explanation for this appears to be the periodic nature of the IRN stimulus itself. Although the perceived pitch of the IRN stimulus was held constant in the case of the steady-state condition, it nevertheless contains a degree of amplitude modulation (visible as dark vertical bands in **Figure 1A**, center), and this is itself a rapid temporal feature. Similarly, the brief nature of the auditory stimuli (250 ms) yield a rapid rise and fall in amplitude envelope during stimulus presentation. We suggest that either of these amplitude modulation characteristics would tend to drive greater response in left vs. right auditory cortex due to their rapid temporal nature. Notwithstanding, this bias cannot fully explain the greater extent of left hemisphere activation in the second univariate analysis, where FM sweeps were contrasted with steady state stimuli. Here again, the left-hemisphere preference persisted even though amplitude modulation features were held constant between the steady-state and FM stimuli.

We next examined whether either the rate or direction of FM sweeps elicited differences in the extent and/or magnitude of activation within auditory cortex. Two contrasts were performed, for the rate and direction of frequency modulation. Neither yielded significant differences, even when a more lenient statistical approach was adopted that allowed for smaller extents of significant voxels. The null findings are not surprising considering what is currently known about the auditory cortex as it pertains to processing FM, the response properties of neurons located in this region, as well as their organization on a macroscopic level. For

instance, using a similar univariate approach, Hsieh et al. (2012) also failed to observe macroscopic regions that differentiated either the direction or rate of FM sweeps. And although electrophysiological work in animals (Mendelson et al., 1993; Tian and Rauschecker, 2004) has revealed the existence of rate-selective and direction-selective neurons for rapid temporal FM stimuli, the selectivity of these neurons is not strictly exclusive. Though some neurons appear to respond more strongly to a specific rate or direction, they also fire at lower levels for other stimulus types as well. Moreover, such neurons are not distributed in a topographically consistent manner, such that a neuron sensitive to one direction or rate might be located immediately adjacent to a neuron sensitive to different parameters. The coarse resolution of fMRI means it would be rather difficult to capture such effects using traditional univariate approaches.

MULTIVARIATE ANALYSIS IDENTIFIES FM-SENSITIVE REGIONS

To address this, RSA was used to perform MVPA analyses in left and right hemisphere auditory cortex. This analysis approach is especially adept at detecting differences in stimulus-dependent patterns of brain activity in the absence of differences in either the magnitude or location of activation. We first investigated whether the direction of frequency-modulated IRN sweeps elicited differentiable patterns of brain activity. We found significant dissimilarity in the patterns of activation for rising vs. falling sweeps bilaterally. This effect was strongest for the broader Auditory cortex ROI, compared to the more narrowly proscribed Heschl's gyrus, suggesting that portions of the belt region outside primary auditory cortex are tuned to FM features of sounds. Indeed, the cytoarchitectonic organization of neurons within core and belt regions of auditory cortex varies considerably and this has implications for the types of analyses that will prove useful in identifying differences in brain activations in response to different acoustic stimuli. While neurons within the auditory core are comprised of smaller, more densely packed neurons, the belt regions that surround it consist of larger and less densely packed neurons (Sweet et al., 2005). These neuroanatomical divisions might serve to drive differences in the representational capacity of these different regions for certain types of acoustic features.

One caveat is in order here: the direction of modulation was manipulated by modifying the initial frequency of the tone sweep. As a result the falling stimulus necessarily had a higher initial frequency than the rising stimulus. Because of this, it is possible that differences between rising and falling stimuli were due to these spectral differences, rather than their FM characteristics. Note that this confound represented what we felt was the least problematic of different possible ways to manipulate direction of frequency modulation; the alternative would have been to create sweeps that involve frequency modulations with different initial and final frequencies such that the overall frequency range of the sweeps was identical but in opposing directions. However, this would have required having a different final steady-state frequency component for rising and falling stimuli (cf. **Table 1**), which because of its duration would have yielded a much stronger spectral confound than what was found here. We do note that our findings are convergent with what Hsieh et al. (2012) found for direction-sensitivity however; their study also manipulated FM direction but controlled for the overall frequency range of both stimulus types by using different stimulus durations. The fact that both our studies have identified direction-sensitive patterns of activation in auditory cortex supports the interpretation that these effects are due to temporal, and not spectral, characteristics of the stimuli.

FAILURE TO IDENTIFY EFFECTS OF FM RATE AT THIS TIME SCALE

Notably, we failed to find similar evidence of sensitivity to the FM rate manipulation in our experiment. This is surprising given a prior affirmative finding by Hsieh et al. (2012) for slower-rate FM sweeps. We argue that the reason for this discrepancy is the short, rapid FM sweep contrasts used in the present study. A study by Schwab et al. (1981) seems especially relevant in this regard. The authors examined adult English speakers' sensitivity to the duration, rate and extent (i.e., frequency range) of a formant transition cue, in the context of discriminating the syllables /ba/ and /wa/. The acoustic characteristics of these formant transitions were very similar to the non-speech FM stimuli used in the present study. The authors found that listeners were sensitive to both the duration and the extent of a formant transition cue; however the rate of frequency change alone was not sufficient for discriminating among phoneme categories. Listeners also appeared to label the two stimuli based on a criterion that weighted both extent and duration equally, such that if the frequency extent (Hz) times the duration (ms) exceeded 23,000, it would be labeled as a glide (/w/), otherwise it would be labeled as a stop (/b/). Note that if we use the same metric for our stimuli, both the "fast" and "slow" rates would fall on the high side of this criterion, due to the relatively narrow frequency extent that we used here (600–1200 Hz).

Overall then, the null result for rate could be interpreted as showing that the phonetic labeling criterion identified by Schwab et al. is in fact recapitulated by the cortical organization within the auditory system. Ultimately however, it would be important in future work to manipulate the extent and duration of FM information in a way that better captures the use of those parameters in phoneme contrasts.

The fact that Hsieh et al. (2012) did find an effect of FM rate appears to also be due to the acoustic parameters that were being used in that study. As noted above, their FM rate manipulation was on a generally slower order than what was used here, and was more in keeping with tonal contrasts observed in some languages. Thus, we are not claiming that modulation rate is never important for speech perception, or that auditory cortex is generally insensitive to such cues. Rather we argue that this is a relatively weak cue at the rapid time scale being considered in this study. Put another way, temporal cues are argued to be relevant to speech cues at multiple grain sizes including phonemes, tones, word-level stress and sentence level stress (Giraud and Poeppel, 2012; Henry and Obleser, 2013). It appears, however, that the types of temporal cues used for different levels of processing may be distinct, and indeed governed by somewhat different principles of neural processing (Obleser et al., 2012).

That said, there is an alternative possibility for our failure to find an effect of sweep rate, which is that our methodology was not sufficiently sensitive to observe such a difference. We used a

rapid, jittered, event-related fMRI paradigm that optimized the ability to present single trials in random order. Our motivation to adopt this design over a block design was that this second option involves repetitive presentations of a given stimulus category within each block, which can inadvertently direct subjects' attention toward the feature of interest for that block. This in turn could yield undesirable effects given our goal of measuring basic perception of acoustic features in auditory cortex. Additionally, the periodic presentation of stimulus trains within each block is itself a temporal feature (i.e., the rise and fall of an amplitude envelope), and this might also drive auditory temporal processes that are separate from the single-stimulus properties that were of interest in our study. Thus, we felt an event-related paradigm, especially one that presented stimuli at irregular intervals, would yield the clearest picture with respect to basic auditory cortical sensitivity to frequency modulation.

On the other hand, it is well recognized that block designs generally yield better statistical power than event-related designs by maximizing the contrast of task-driven BOLD response against background noise. It is therefore conceivable that we would have found effects of sweep rate had we adopted a block design. We do note however that there was sufficient power in our experiment to find effects of sweep direction using the same analyses. Given that the same number of trials was employed for both manipulations, we can at the very least conclude that the effect of rapid FM direction is appreciably stronger than that of FM rate.

IMPLICATIONS FOR PHONETIC PROCESSING

Recognizing speech extends beyond just recognizing the component acoustic features of the speech stream. What we have examined here is an early step in a processing chain that involves matching the acoustic features to phonetic categories and/or articulatory gestures, and proceeding onwards to lexical, semantic and syntactic analyses (Hickok and Poeppel, 2004; Rauschecker and Scott, 2009). So for example, other studies have found that phonetic perception, in which speech sounds are categorized or discriminated, specifically engages STS areas that are ventral to the STG regions of interest in this study (Liebenthal et al., 2005; Joanisse et al., 2007). Likewise, sounds that are perceived as speech yield fMRI effects that are differentiable from those observed for acoustically similar non-speech sounds, again supporting the view that the phonetic content of speech engages selective brain mechanisms beyond simple acoustic feature detection (Vouloumanos et al., 2001; Liebenthal et al., 2005). So in short, what we have identified in the present study might be best thought of as the acoustic precursors to acoustic-phonetic perception, and cannot explain the entire process of phoneme recognition during speech perception. We do predict however that speech sounds that contain similar acoustic cues to the non-speech cues manipulated here will also yield similar effects in the regions of auditory cortex, supporting the view that at an early point in processing there is no strong distinction between how speech and non-speech sounds are processed.

CONCLUSIONS

We used fMRI to examine the organization of human auditory cortex for processing frequency modulated sounds. The results yield insights into how auditory cortex processes acoustic elements that are fundamental to phoneme perception. Using IRN stimuli that approximate both spectral and rapid temporal speech characteristics, we observed that FM sweeps activated a broader set of regions of auditory cortex compared to control sounds that were spectrally similar but not frequency-modulated. More importantly, multivariate analyses demonstrated the existence of direction-specific activity patterns at a microscopic level in both left and right auditory cortex. The findings add to a growing literature supporting the view that auditory cortex contains neural populations specifically tuned to detecting at least some types of acoustic features important for phonetic processing. Moreover it illustrates the utility of applying multivariate data analysis techniques such as RSA to elucidate differences in patterns of brain activity when gross regions of activation overlap.

ACKNOWLEDGMENTS

This research was funded by a Discovery Grant and Accelerator Award to MFJ from the Natural Sciences and Engineering Research Council (Canada). We are deeply grateful for the assistance of Annika Linke, Chris McNorgan and Connor Wild for providing the ROIs used in this study and assisting with RSA analyses. Thanks to Jackson Gandour and Jayaganesh Swaminathan for providing the Matlab code for generating IRN stimuli.

REFERENCES

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

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc. Inst. Phon. Sci.* 17, 97–110.

Carrasco, A., and Lomber, S. G. (2011). Neuronal activation times to simple, complex, and natural sounds in cat primary and non-primary auditory cortex. *J. Neurophysiol.* 106, 1166–1178. doi: 10.1152/jn.00940.2010

Chevillet, M., Riesenhuber, M., and Rauschecker, J. P. (2011). Functional correlates of the anterolateral processing hierarchy in human auditory cortex. *J. Neurosci.* 31, 9345–9352. doi: 10.1523/JNEUROSCI.1448-11.2011

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014

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

Godey, B., Atencio, C. A., Bonham, B. H., Schreiner, C. E., and Cheung, S. W. (2005). Functional organization of squirrel monkey primary auditory cortex: responses to frequency-modulation sweeps. *J. Neurophysiol.* 94, 1299–1311. doi: 10.1152/jn.00950.2004

Hall, D. A., Johnsrude, I. S., Haggard, M. P., Palmer, A. R., Akeroyd, M. A., and Summerfield, A. Q. (2002). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 11, 946–953. doi: 10.1093/cercor/12.2.140

Heinemann, L. V., Rahm, B., Kaiser, J., Gaese, B. H., and Altmann, C. F. (2010). Repetition enhancement for frequency-modulated but not unmodulated sounds: a human MEG study. *PLoS ONE* 5:e15548. doi: 10.1371/journal.pone.0015548

Henry, M. J., and Obleser, J. (2013). Dissociable neural response signatures for slow amplitude and frequency modulation in human auditory cortex. *PLoS ONE* 8:e78758. doi: 10.1371/journal.pone.0078758

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111. doi: 10.1121/1.411872

Hsieh, I. H., Fillmore, P., Rong, F., Hickok, G., and Saberi, K. (2012). FM-selective networks in human auditory cortex revealed using fMRI and multivariate pattern classification. *J. Cogn. Neurosci.* 24, 1896–1907. doi: 10.1162/jocn_a_00254

Joanisse, M. F., and Gati, J. S. (2003). Overlapping neural regions for processing rapid temporal cues in speech and nonspeech signals. *Neuroimage* 19, 64–79. doi: 10.1016/S1053-8119(03)00046-6

Joanisse, M. F., Zevin, J. D., and McCandliss, B. D. (2007). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. *Cereb. Cortex* 17, 2084–2093. doi: 10.1093/cercor/bhl124

Kaas, J. H., Hackett, T. A., and Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Curr. Opin. Neurobiol.* 9, 164–170. doi: 10.1016/S0959-4388(99)80022-1

Kikuchi, Y., Horwitz, B., and Mishkin, M. (2010). Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J. Neurosci.* 30, 13021–13030. doi: 10.1523/JNEUROSCI.2267-10.2010

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 1–28. doi: 10.3389/neuro.06.004.2008

Kusmierek, P., and Rauschecker, J. P. (2009). Functional specialization of medial auditory belt cortex in the alert rhesus monkey. *J. Neurophysiol.* 102, 1606–1622. doi: 10.1152/jn.00167.2009

Lewis, J. W., Brefczynski, J. A., Phinney, R. E., Janik, J. J., and DeYoe, E. A. (2005). Distinct cortical pathways for processing tool versus animal sounds. *J. Neurosci.* 25, 5148–5158. doi: 10.1523/JNEUROSCI.0419-05.2005

Lewis, J. W., Talkington, W. J., Walker, N. A., Spirou, G. A., Jajosky, A., Frum, C., et al. (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. *J. Neurosci.* 29, 2283–2296. doi: 10.1523/JNEUROSCI.4145-08.2009

Liang, L., Lu, T., and Wang, X. (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *J. Neurophysiol.* 87, 2237–2261.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., and Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *J. Exp. Psychol.* 52, 127–137. doi: 10.1037/h0041240

Liebenthal, E., Binder, J. R., Piorkowski, R. L., and Remez, R. E. (2003). Short-term reorganization of auditory analysis induced by phonetic experience. *J. Cogn. Neurosci.* 15, 549–558. doi: 10.1162/089892903321662930

Liebenthal, E., Binder, J. R., and Spitzer, S. M. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Linke, A. C., Vicente-Grabovetsky, A., and Cusack, R. (2011). Stimulus-specific suppression preserves information in auditory short-term memory. *Proc. Natl. Acad. Sci.* 108, 12961–12966. doi: 10.1073/pnas.1102118108

Mendelson, J. R., Schreiner, C. E., Sutter, M. L., and Grasse, K. L. (1993). Functional topography of cat primary auditory cortex: responses to frequency-modulated sweeps. *Exp. Brain Res.* 94, 65–87. doi: 10.1007/BF00230471

Miller, J. L., and Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Percept. Psychophys.* 25, 457–465. doi: 10.3758/BF03213823

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., et al. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *Neuroimage* 30, 563–569. doi: 10.1016/j.neuroimage.2005.10.002

Nelken, I., Fishbach, A., Las, L., Ulanovsky, N., and Farkas, D. (2003). Primary auditory cortex of cats: feature detection or something else? *Biol. Cybern.* 89, 397–406. doi: 10.1007/s00422-003-0445-3

Nelken, I., and Versnel, H. (2000). Responses to linear and logarithmic frequency-modulated sweeps in ferret primary auditory cortex. *Eur. J. Neurosci.* 12, 549–562. doi: 10.1046/j.1460-9568.2000.00935.x

Obleser, J., Herrmann, B., and Henry, M. J. (2012). Neural oscillations in speech: don't be enslaved by the envelope. *Front. Hum. Neurosci.* 6:250. doi: 10.3389/fnhum.2012.00250

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

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111–114. doi: 10.1126/science.7701330

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–949. doi: 10.1126/science.7233191

Schwab, E. C., Sawusch, J. R., and Nusbaum, H. C. (1981). The role of second formant transitions in the stop-semivowel distinction. *Percept. Psychophys.* 29, 121–128. doi: 10.3758/BF03207275

Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123, 2400–2406. doi: 10.1093/brain/123.12.2400

Stevens, K. N., and Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Am.* 55, 653–659. doi: 10.1121/1.1914578

Swaminathan, J., Krishnan, A., Gandour, J. T., and Xu, Y. (2008). Applications of static and dynamic iterated rippled noise to evaluate pitch encoding in the human auditory brainstem. *IEEE Trans. Biomed. Eng.* 50, 281–287. doi: 10.1109/TBME.2007.896592

Sweet, R. A., Dorph-Petersen, K. A., and Lewis, D. A. (2005). Mapping auditory core, lateral belt, and parabelt cortices in the human superior temporal gyrus. *J. Comput. Neurol.* 491, 270–289. doi: 10.1002/cne.20702

Talairach, J., and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging*. New York, NY: Thieme Medical.

Tian, B., and Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* 92, 2993–3013. doi: 10.1152/jn.00472.2003

Vouloumanos, A., Kiehl, K. A., Werker, J. F., and Liddle, P. F. (2001). Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to speech and nonspeech. *J. Cogn. Neurosci.* 13, 994–1005. doi: 10.1162/089892901753165890

Washington, S. D., and Kanwal, J. S. (2008). DSCF neurons within the primary auditory cortex of the mustached bat process frequency modulations present within social calls. *J. Neurophysiol.* 100, 3285–3304. doi: 10.1152/jn.90442.2008

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., and Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *J. Cogn. Neurosci.* 13, 1–7. doi: 10.1162/089892901564108

Whalen, D. H., Benson, R. R., Richardson, M., Swainson, B., Clark, V. P., Lai, S., et al. (2006). Differentiation of speech and nonspeech processing within primary auditory cortex. *J. Acoust. Soc. Am.* 119, 575–5814. doi: 10.1121/1.2139627

Zatorre, R. J., and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 11, 946–953. doi: 10.1093/cercor/11.10.946

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

Received: 28 March 2014; accepted: 11 September 2014; published online: 30 September 2014.

*Citation: Joanisse MF and DeSouza DD (2014) Sensitivity of human auditory cortex to rapid frequency modulation revealed by multivariate representational similarity analysis. Front. Neurosci. **8**:306. doi: 10.3389/fnins.2014.00306*

This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 December 2014
doi: 10.3389/fnins.2014.00406

Hierarchical organization of speech perception in human auditory cortex

Colin Humphries[1]*, Merav Sabri[1], Kimberly Lewis[1] and Einat Liebenthal[1,2]

[1] *Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA*
[2] *Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, USA*

Edited by:
Marc Schönwiesner, University of Montreal, Canada

Reviewed by:
Teemu Rinne, University of Helsinki and Aalto University School of Science, Finland
David L. Woods, University of California, Davis/VANCHCES, USA

****Correspondence:***
Colin Humphries, Department of Neurology, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226, USA
e-mail: chumphri@mcw.edu

Human speech consists of a variety of articulated sounds that vary dynamically in spectral composition. We investigated the neural activity associated with the perception of two types of speech segments: (a) the period of rapid spectral transition occurring at the beginning of a stop-consonant vowel (CV) syllable and (b) the subsequent spectral steady-state period occurring during the vowel segment of the syllable. Functional magnetic resonance imaging (fMRI) was recorded while subjects listened to series of synthesized CV syllables and non-phonemic control sounds. Adaptation to specific sound features was measured by varying either the transition or steady-state periods of the synthesized sounds. Two spatially distinct brain areas in the superior temporal cortex were found that were sensitive to either the type of adaptation or the type of stimulus. In a relatively large section of the bilateral dorsal superior temporal gyrus (STG), activity varied as a function of adaptation type regardless of whether the stimuli were phonemic or non-phonemic. Immediately adjacent to this region in a more limited area of the ventral STG, increased activity was observed for phonemic trials compared to non-phonemic trials, however, no adaptation effects were found. In addition, a third area in the bilateral medial superior temporal plane showed increased activity to non-phonemic compared to phonemic sounds. The results suggest a multi-stage hierarchical stream for speech sound processing extending ventrolaterally from the superior temporal plane to the superior temporal sulcus. At successive stages in this hierarchy, neurons code for increasingly more complex spectrotemporal features. At the same time, these representations become more abstracted from the original acoustic form of the sound.

Keywords: speech perception, auditory cortex, phonological processing, fMRI, temporal lobe, spectrotemporal cues

INTRODUCTION

During the articulation of speech, vibrations of the vocal cords create discrete bands of high acoustic energy called formants that correspond to the resonant frequencies of the vocal tract. Identifying phonemic information from a speech stream depends on both the steady-state spectral content of the sound, particularly the relative frequencies of the formants, and the temporal content, corresponding to fast changes in the formants over time. Speech sounds can be divided into two general categories, vowels and consonants, depending on whether the vocal tract is open or obstructed during articulation. Because of this difference in production, vowels, and consonants have systematic differences in acoustic features. Vowels, which are produced with an open vocal tract, generally consist of sustained periods of sound with relatively little variation in frequency. Consonants, on the other hand, are voiced with an obstructed vocal tract, which tends to create abrupt changes in the formant frequencies. For this reason, vowel identification relies more heavily on the steady-state spectral features of the sound and consonant identification relies more on the momentary temporal features (Kent, 2002).

Research in animals suggests that the majority of neurons in auditory cortex encode information about both spectral and temporal properties of sounds (Nelken et al., 2003; Wang et al., 2008; Bendor et al., 2012). However, the spectrotemporal response properties of neurons vary across cortical fields. For example, in the core region of primate auditory cortex, neurons in anterior area R integrate over longer time windows than neurons in area A1 (Bendor and Wang, 2008; Scott et al., 2011), and neurons in the lateral belt have preferential tuning to sounds with wide spectral bandwidths compared to the more narrowly-tuned neurons in the core (Rauschecker et al., 1995; Rauschecker and Tian, 2004; Recanzone, 2008). This pattern of responses has been used as evidence for the existence of two orthogonal hierarchical processing streams in auditory cortex: a stream with increasing longer temporal windows extending along the posterior-anterior axis from A1 to R and a stream with increasing larger spectral bandwidth extending along the medial-lateral axis from the core to the belt (Rauschecker et al., 1995; Bendor and Wang, 2008). In addition to differences in spectrotemporal response properties within auditory cortex, other studies suggest there may also be differences between the two hemispheres, with the right hemisphere more

sensitive to fine spectral details and the left hemisphere more sensitive to fast temporal changes (Zatorre et al., 2002; Poeppel, 2003; Boemio et al., 2005).

In the current study functional magnetic resonance imaging (fMRI) was used to investigate the cortical organization of phonetic feature encoding in the human brain. A main question is whether there are spatially distinct parts of auditory cortex that encode information about spectrally steady-state and dynamic sound features. Isolating feature-specific neural activity is often a problem in fMRI because different features of a stimulus may be encoded by highly overlapping sets of neurons, which could potentially result in similar patterns and levels of BOLD activation during experimental manipulations. One way to improve the sensitivity of fMRI to feature-specific encoding is to use stimulus adaptation (Grill-Spector and Malach, 2001). Adaptation paradigms rely on the fact that neural activity is reduced when a stimulus is repeated, and this effect depends on the type of information the neuron encodes. For example, a visual neuron that encodes information about spatial location might show reduced activity when multiple stimuli were presented in the same location, but would be insensitive to repetition of other features like color or shape. Adaptation-type paradigms have been used previously to study aspects of speech processing, such as phonemic categorization (Wolmetz et al., 2010), consonant (Lawyer and Corina, 2014), and vowel processing (Leff et al., 2009). In the current study, subjects listened to stimuli that were synthetic two-formant consonant-vowel (CV) syllables composed of an initial period of fast temporal change, corresponding primarily to the consonant, and a subsequent steady-state period, corresponding to the vowel. These stimuli were presented in an adaptation design, in which each trial consisted of a series of four identical syllables (e.g., /ba/, /ba/, /ba/, /ba/) followed by two stimuli that differed either in the initial transition period (e.g,. /ga/, /ga/), the steady-state period (e.g., /bi/, /bi/), or both (e.g., /gi/, /gi/). A fourth condition, in which all six stimuli were identical, was included as a baseline. The baseline condition should produce the greatest amount of stimulus adaptation and the lowest activation levels. We expected that trials with changes in the transition period compared to baseline trials would result in greater activity in neurons that encode information about fast temporal transitions, while trials with changes in the steady-state period would result in greater activity in neurons that encode information about spectral composition.

An additional question is whether any observed activation patterns represent differences in general auditory processing or differences specific to the processing of speech vowels and consonants. Previous imaging studies comparing activation during consonant and vowel processing have only used speech stimuli (Rimol et al., 2005; Obleser et al., 2010) or have used non-speech controls that were acoustically very different from speech (Joanisse and Gati, 2003), making it difficult to determine speech specificity. To address this question, we included two types of acoustically matched non-phonemic control sounds. In one type, the first formant was spectrally rotated, resulting in a sound with the same spectral complexity of speech but including a non-native (in English) formant transition. The second type of control stimuli included only one of the formants, resulting in a sound with valid English formant transitions but without harmonic spectral content. These three stimulus types (phonemic, non-phonemic, single-formant) were presented in trials of six ordered according to the four types of adaptation (steady-state change, transition change, steady-state and transition change, baseline) resulting in 12 conditions.

MATERIALS AND METHODS

PARTICIPANTS

FMRI data were collected from 15 subjects (8 female, 7 male; ages 21–36 years). All subjects were right-handed, native English speakers, and had normal hearing based on self report. Subjects gave informed consent under a protocol approved by the Institutional Review Board of the Medical College of Wisconsin.

STIMULI

The stimuli were synthesized speech sounds created using the KlattGrid synthesizer in Praat (http://www.fon.hum.uva.nl/praat). The acoustic parameters for the synthesizer were derived from a library of spoken CV syllables based on a male voice (Stephens and Holt, 2011). For each syllable, we first estimated the center frequencies of the first and second formants using linear predictive coding (LPC). Outliers in the formant estimates were removed. The timing of the formant estimates were adjusted so that the duration of the initial transition period of each syllable was 40 ms and the duration of the following steady-state period was 140 ms. The resulting formant time series were used as input parameters to the speech synthesizer. Three types of stimuli were generated (see **Figure 1A**). Phonemic stimuli were composed of both the F1 and F2 formant time courses derived from the natural syllables. Non-Phonemic stimuli were composed of the same F2 formants as the Phonemic stimuli and a spectrally rotated version of the F1 formant (inverted around the mean frequency of the steady-state period). Single-Formant stimuli contained only the F1 or F2 formant from the Phonemic and Non-Phonemic stimuli. Qualitatively, the Phonemic stimuli were perceived as English speech syllables, the Non-Phonemic stimuli were perceived as unrecognizable (non-English) speech-like sounds, and the Single-Formant stimuli were perceived as non-speech chirps (Liebenthal et al., 2005). Versions of these three types of synthesized stimuli were generated using all possible combinations of the consonants /b/, /g/, /d/, and the vowels /a/, /ae/, /i/, and /u/. Perception of the resulting stimuli was then tested in a pilot study, in which subjects ($n = 6$) were asked to identify each stimulus as one of the 12 possible CV syllables, as a different CV syllable, or as a non-speech sound. Based on the pilot study results, several of the Non-Phonemic and Single-Formant stimuli were removed from the stimulus set because they sounded too speech-like, and several of the Phonemic stimuli were removed because they were too often misidentified for another syllable or non-speech sound. A final stimulus set was chosen that consisted of Phonemic, Non-Phonemic, and Single-Formant versions of the syllables: /ba/, /bi/, /bae/, /ga/, /gi/, /gae/. In the final set, the Phonemic, Non-Phonemic, and Single-Formant stimuli were identified by participants of the pilot study as the original syllable (from which the syllable was derived and re-synthesized) at an average accuracy of 90, 46, and 13%, respectively.

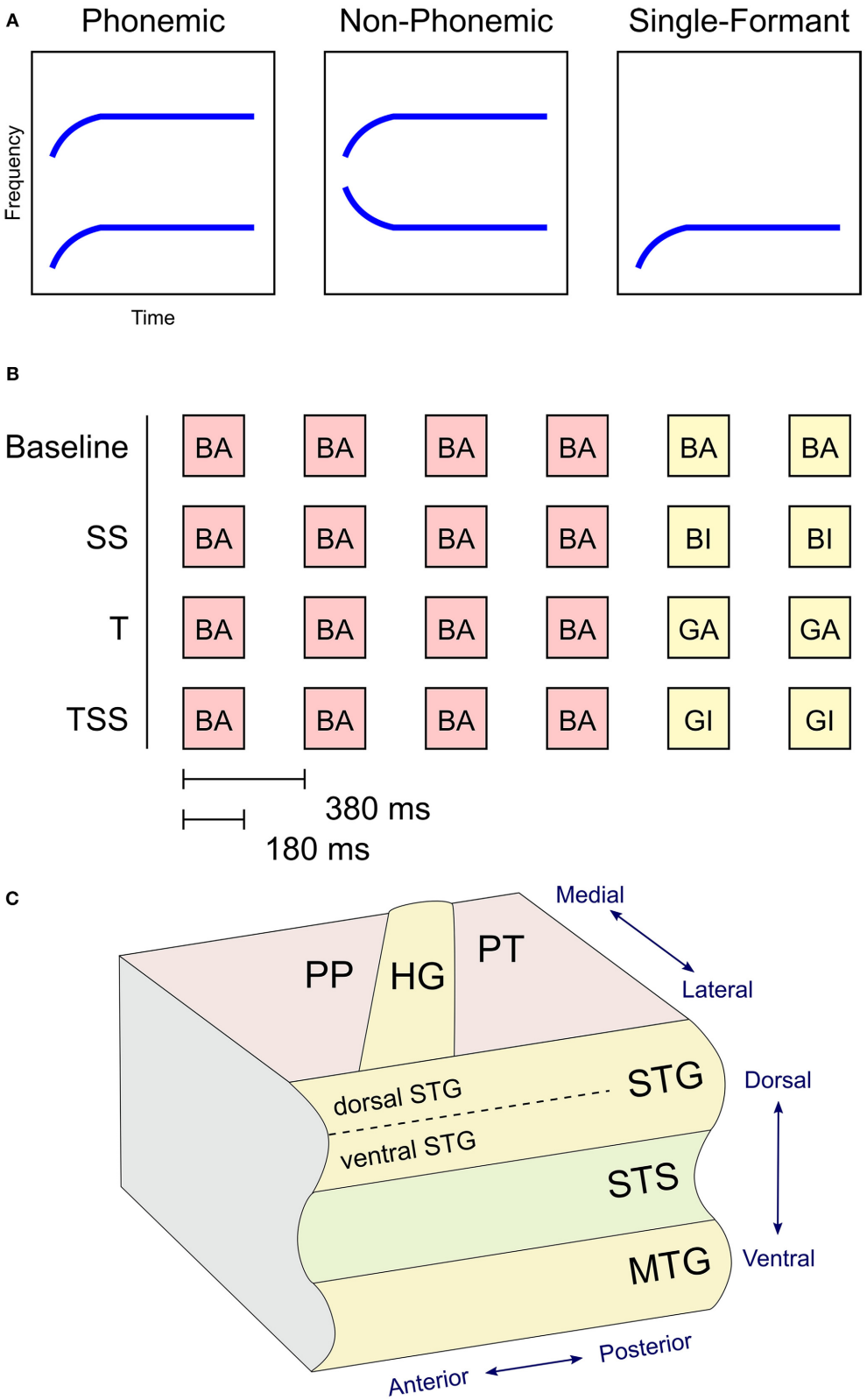


FIGURE 1 | (A) Stimulus design. Graphs illustrate the shape of the formants used to synthesize the three types of stimuli based on the syllable /ba/. Phonemic stimuli were synthesized using the first (F1) and second (F2) formants in their canonical orientation. Non-Phonemic stimuli were composed of a standard F2 formant and a spectrally rotated F1 formant. Single-Formant stimuli only included one of the two formants (F1 or F2) from the Phonemic or Non-Phonemic stimuli. **(B)** Trial design. Examples of the four

(Continued)

FIGURE 1 | Continued
adaptation conditions are shown. Each trial consisted of six stimuli presented every 380 ms. The first four stimuli were identical. The last two stimuli varied in one of four ways. In Baseline trials the final two stimuli were identical to the first four. In Steady-State (SS) trials, the final two stimuli differed in the steady-state period (i.e., vowel). In Transient (T) trials, the final two stimuli different in the initial transition period (i.e., consonant). In the Transient and Steady-State (TSS) trials both transient and steady-state periods differed in the final two stimuli. **(C)** Diagram of superior and middle temporal cortex in the left hemisphere with labeled anatomical structures. Abbreviations: PP, Planum Polare; PT, Planum Temporale; HG, Heschl's Gyrus; STG, Superior Temporal Gyrus; STS, Superior Temporal Sulcus; MTG, Middle Temporal Gyrus.

The stimuli were presented using an adaptation paradigm (see **Figure 1B**). Each trial contained six stimuli presented every 380 ms. The first four stimuli were identical, and the final two stimuli differed from the first four in one of four ways. In the Baseline condition, the final two stimuli were identical to the first four. In the Steady-State (SS) condition, the final two stimuli differed from the first four in the steady-state vowel (e.g., /ba/, /ba/, /ba/, /ba/, /bi/, /bi/). In the Transition (T) condition, the final stimuli differed in their transition period (e.g., /ba/, /ba/, /ba/, /ba/, /ga/, /ga/). In the Transition Steady-State (TSS) condition, both the steady-state and transition periods differed in the final stimuli (e.g., /ba/, /ba/, /ba/, /ba/, /gi/, /gi/).

PROCEDURE

Each participant was scanned in two sessions occurring on different days. Each scanning session consisted of a high resolution anatomical scan (SPGR sequence, axial orientation, 180 slices, 256×240 matrix, FOV = 240 mm, $0.9375 \times 1.0\,mm^2$ resolution, 1.0 mm slice thickness) and five functional scans (EPI sequence, 96×96 matrix, FOV = 240 mm, $2.5 \times 2.5\,mm^2$ resolution, 3 mm slice thickness, $TA = 1.8$ s, $TR = 7.0$ s). Functional scans were collected using a sparse-sampling procedure in which stimuli were presented during a silent period between MR image collection (Hall et al., 1999).

The experiment was organized in a 3×4 factorial design with the three stimulus types (Phonemic, Non-Phonemic, and Single-Formant) presented in four different adaptation configurations (TSS, T, SS, and Control) resulting in a total of 12 conditions. The conditions were presented in trials consisting of six stimuli presented every 380 ms followed by a single MR volume acquisition lasting 1.8 s. A small percentage ($p = 0.1$) of trials were missing either one or two of the six stimuli. To ensure that subjects were attending to the stimuli during the experiment, subjects were required to hit a button when they detected a missing stimulus. Compliance with the task was assessed, but image data from the trials with missing stimuli were excluded from the analysis. Within each run 8 trials were presented per condition producing a total of 80 trials per condition across both sessions. An additional 8 trials of rest (i.e., no stimulus) were included in each run. Trials were presented in blocks containing 4 trials of the same condition. The order of the blocks was randomized across runs and across participants.

Sounds were presented binaurally with in-ear electrostatic headphones (Stax SR-003; Stax Ltd, Saitama, Japan). Additional protective ear muffs were placed over the headphones to attenuate scanner noise.

The fMRI data were analyzed using AFNI (Saad et al., 2009). Initial preprocessing steps included motion correction and co-registration between the functional and anatomical scans. The anatomical volumes from each subject were aligned using non-linear deformation to create a study-specific atlas using the program ANTS (Avants and Gee, 2004). The functional data were resampled (voxel size $= 2.5 \times 2.5 \times 2.5\,mm^3$) into the atlas space and spatially filtered using a Gaussian window (FWHM = 5 mm). Our primary research questions were focused on differences in activation in auditory areas, therefore, we confined our analysis to a set of voxels that included the entire superior, middle, and inferior temporal lobe and extending into the inferior parietal and lateral occipital lobes.

Estimates of the activation levels for the 12 conditions were calculated using the AFNI command 3dREMLfit, which models the data using a generalized least squares analysis with a restricted maximum likelihood (REML) estimate of temporal auto-correlation. Contrasts between conditions were evaluated at the group level using a mixed-effects model. To correct for increased type 1 error due to multiple comparisons, the voxels in the resulting statistical maps were initially thresholded at $p < 0.01$, grouped into contiguous clusters, and then thresholded at $p < 0.05$ using a cluster-size threshold of 29 determined using the AFNI command 3dClustStim. An additional analysis using an initial threshold of $p < 0.05$ and a cluster-size threshold of 108 voxels ($p < 0.05$, corrected) was performed on one of the contrasts. Mean effect sizes for each cluster were calculated by dividing the amplitude of the contrast values by the mean signal level and then taking a mean across all the voxels in the cluster. The maps are displayed on an inflated surface brain of the ANTS-derived atlas created using Freesurfer (Dale et al., 1999). A diagram of the location of the anatomical labels used to describe the results is displayed in **Figure 1C**.

RESULTS

Differences in BOLD activation between the three stimulus types are shown in **Figure 2**. Each contrast represents the difference in activation between two of the three stimulus types collapsed across the four adaptation conditions. Greater levels of activity were observed during Phonemic trials compared to either the Non-Phonemic or Single-Formant trials in the superior temporal gyrus (STG), bilaterally. More specifically, the voxels in this activation cluster were located on the more inferior side of the curve of the STG (see **Figure 4**), which we refer to as ventral STG, and distinguish this area from the more superior side of the STG, which we refer to as dorsal STG. There was less activity during Phonemic trials compared to Single-Formant trials in both hemispheres in the superior temporal plane (STP), specifically the medial portion, and in the posterior part of the middle temporal sulcus. Less activity during Phonemic compared to Non-Phonemic trials was found in a smaller cluster in the planum polare in the right hemisphere. Single-Formant trials

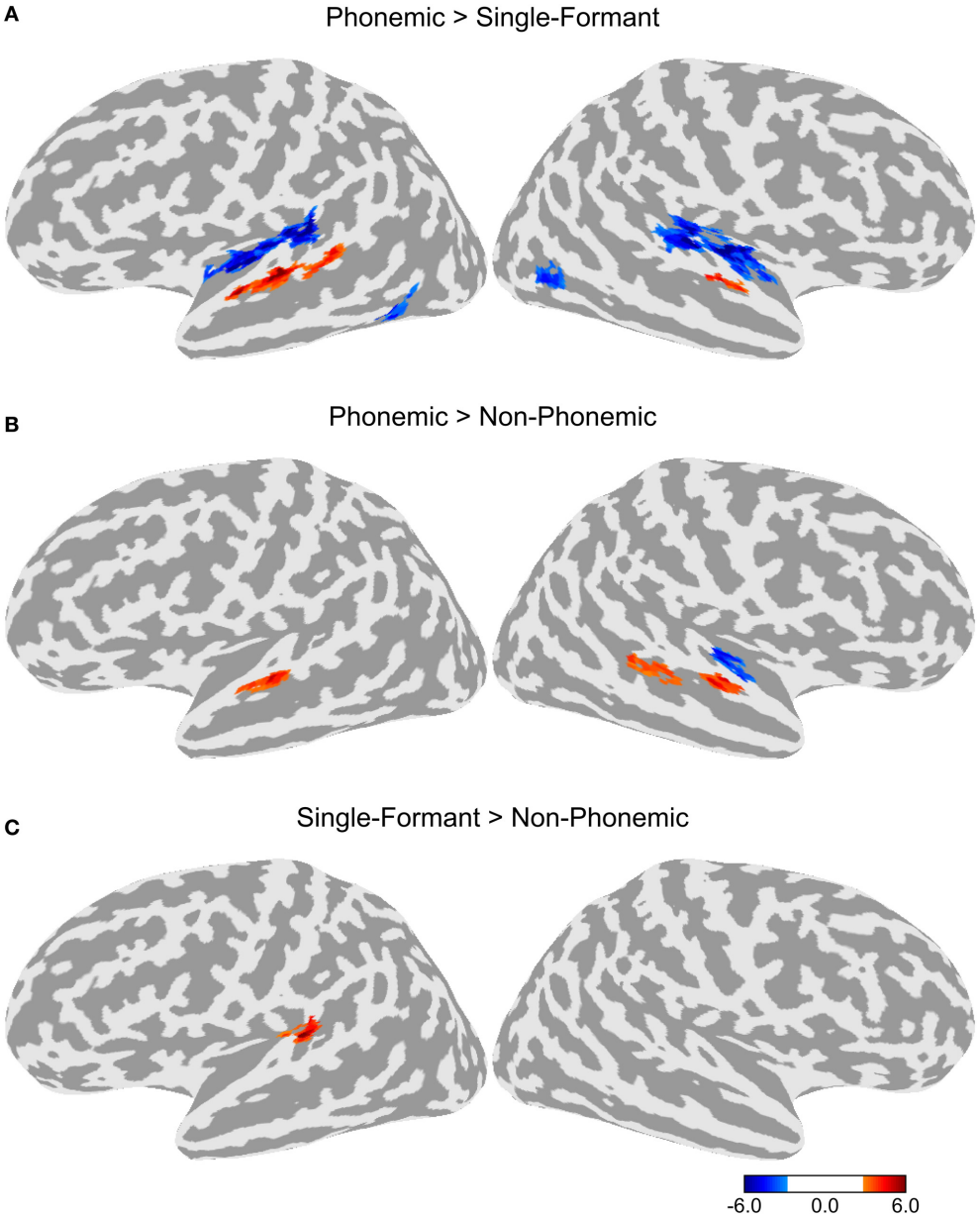


FIGURE 2 | Differences in activation between the three stimulus types collapsed across the four adaptation conditions. (A) Comparison of the activation levels in the Phonemic and Single-Formant trials. **(B)** Comparison between the Phonemic and Non-Phonemic trials. **(C)** Comparison between the Single-Formant and Non-Phonemic trials.

had greater activity than Non-Phonemic trials in the left planum temporale.

To test for adaptation effects, each of the three adaptation conditions (T, SS, and TSS) were compared to the Baseline adaptation condition, in which all six stimuli in the trial were identical. Each of the adaptation contrasts included all three stimulus types. The resulting maps are shown in **Figure 3**. All three adaptation conditions demonstrated greater activity than the Baseline condition in the dorsal STG, bilaterally. The comparison of SS against Baseline produced a cluster of activation extending along the dorsal STG both anterior and posterior to Heschl's gyrus (HG). The TSS condition activated a similar set of areas. The T condition appeared to have the smallest extent of activation confined to a section of cortex along the middle of the STG. Additional adaptation effects were observed outside of auditory cortex. Significant clusters of activation for the T condition were observed in the left middle temporal gyrus (MTG) and bilateral middle temporal sulcus. In addition, activation for the SS was found in the right lateral occipital sulcus.

A direct contrast between the T and SS conditions is shown in **Figure 4A**. Greater activity in the SS condition was observed in a cluster in the left anterior STG and another cluster in the right

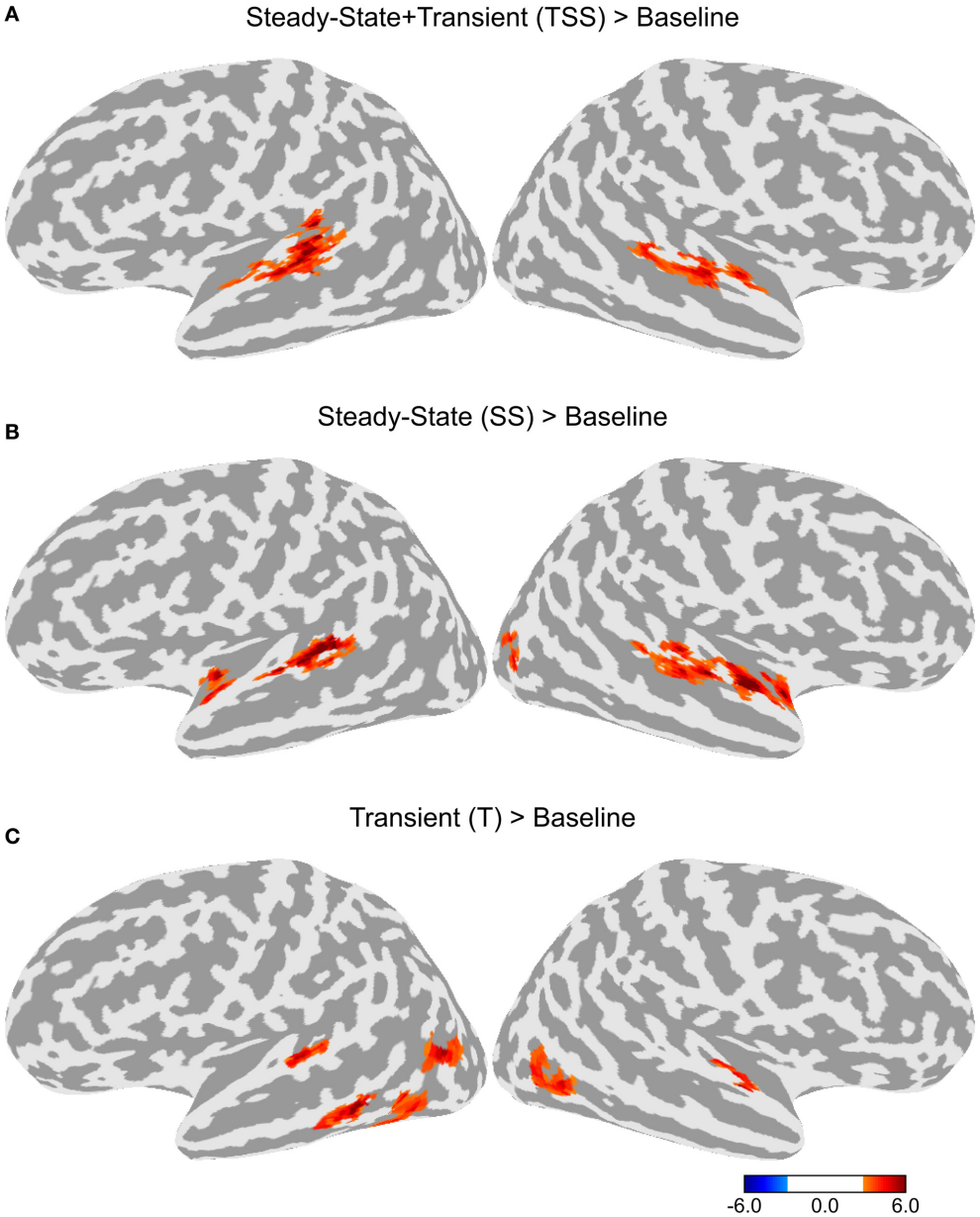


FIGURE 3 | Differences in activation between each adaptation condition and Baseline collapsed across stimulus type. (A) Contrast between activation levels in the Transient and Steady-State (TSS) condition against the Baseline condition. **(B)** Contrast between the Steady-State (SS) condition and the Baseline condition. **(C)** Contrast between the Transient (T) condition and the Baseline condition.

posterior STG. Greater activity in the T condition was observed in the left superior marginal gyrus and the right temporal pole. Given that differences in activation levels between the two types of adaptation could be small resulting in a lower statistical effect, we ran an additional contrast using a lower initial threshold of $p < 0.05$ with the same corrected alpha level of 0.05 (see **Figure 4B**). In this contrast, there was greater activity in the SS condition in bilateral anterior STG and bilateral posterior STG. There was no difference between T and SS in the middle section of the STG just lateral to HG. Greater activation for the T condition was observed in bilateral lateral occipital complex and the left temporal pole.

In order to compare the location of the activation clusters identified in the dorsal and ventral STG, we overlaid the activation maps for the combination of the two stimulus contrasts (Phonemic > Non-Phonemic and Phonemic > Single-Formant) and the three adaptation contrasts (SS > Baseline, T > Baseline, and TSS > Baseline) (**Figure 5**). Voxels that were significant for either of the two stimulus contrasts are displayed in red, voxels significant for any of the three adaptation contrasts are in yellow, and overlapping voxels are in orange. Activation clusters showing preferential response to phonemic stimuli were ventral and adjacent to clusters showing adaptation effects related

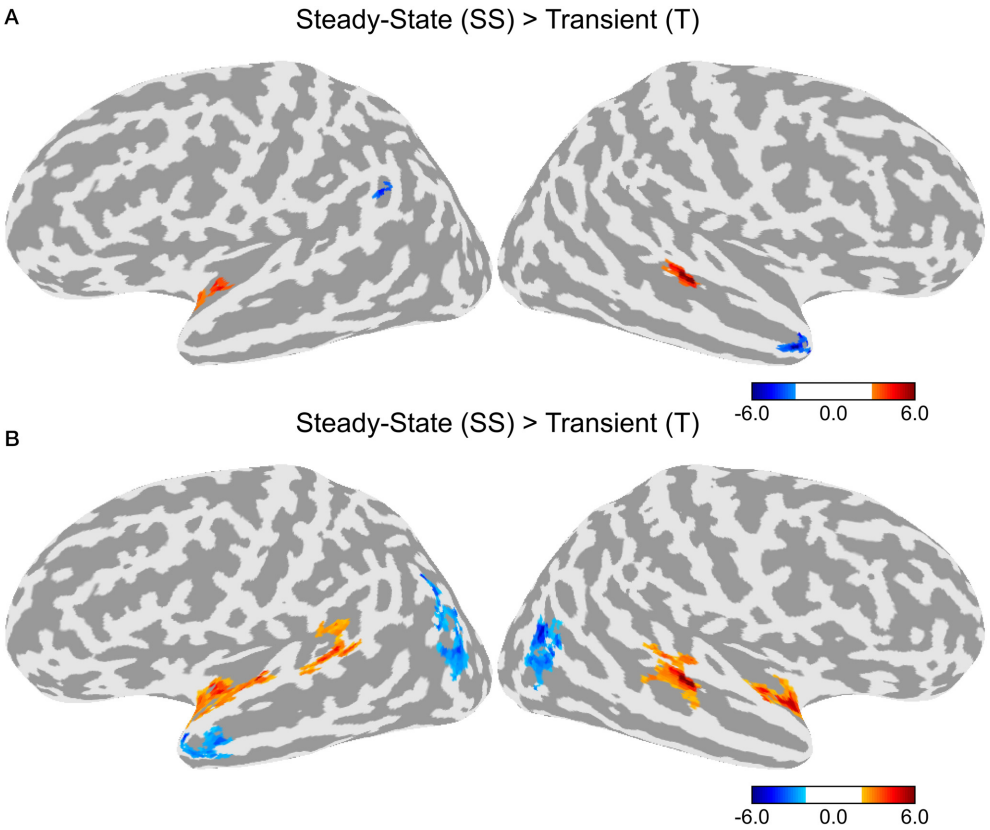


FIGURE 4 | Differences in activation between the Transient (T) and Steady-State (SS) adaptation conditions. (A) Contrast between T and SS using an initial threshold of $p < 0.01$ ($\alpha = 0.05$, corrected). **(B)** Contrast between T and SS using an initial threshold of $p < 0.05$ ($\alpha = 0.05$, corrected).

to changes in acoustic form with little overlap between the clusters.

In the sections of cortex in the dorsal and ventral STG that showed activation in the stimulus and adaptation contrasts, we did not find significant interactions between adaptation and stimulus type. However, significant interaction effects were seen in several clusters outside of this region (see **Table 1**). The interaction between SS and Single-Formant over Phonemic showed a cluster in the right inferior parietal lobe and between SS and Single-Formant over Non-Phonemic in the left middle temporal sulcus. The interaction between T and Phonemic over Single-Formant was seen in the left anterior STS. The interaction between TSS and Phonemic over Non-Phonemic showed activation in the right posterior STS/STG and between TSS and Single-Formant over Non-Phonemic in the bilateral posterior STG and bilateral MTG.

DISCUSSION

We investigated the patterns of neural activity associated with perception of the transition and steady-state portions of CV syllables and non-speech controls using fMRI. Two adjacent but distinct regions in the superior temporal lobe were identified that were affected by manipulations of either feature-specific adaptation or stimulus type (**Figure 5**). On the dorsal side of the STG extending into the STP, voxels had reduced activity during the repetition of both the transition and steady-state portions of the sound regardless of whether the stimulus was Phonemic, Non-Phonemic, or Single-Formant. On the ventral side of the STG extending into the STS, voxels displayed higher levels of activity during Phonemic compared to Non-Phonemic and Single-Formant trials but were not sensitive to adaptation of acoustic features. Brain areas showing selectivity to acoustic form (i.e., to the adaptation condition) and brain areas showing selectivity to phonemes were located adjacent to each other in the dorsal and ventral STG, with little overlap between them. Finally in bilateral STP, increased activity was observed for the Non-Phonemic and Single-Formant sounds over the Phonemic sounds.

Adaptation effects due to stimulus repetition were observed in the bilateral dorsal STG extending into the STP. This region has been identified in a wide range of studies looking at auditory and speech processing (Alho et al., 2014), and it appears to play a role in processing stimuli with "complex" spectrotemporal structure. For example, higher levels of activity in the bilateral dorsal STG are observed for sounds with multiple spectral components (Schönwiesner et al., 2005; Lewis et al., 2012; Moerel et al., 2013; Norman-Haignere et al., 2013) or sounds containing temporal modulations (Schönwiesner et al., 2005; Herdener et al., 2013; Santoro et al., 2014) compared to simple auditory controls like tones or noise. Greater activity is also observed in this

Table 1 | FMRI Activation Clusters.

Hemi	Center			Peak			*t*-value	Cluster size (voxels)	Mean effect size (%)	Region
	X	*Y*	*Z*	*X*	*Y*	*Z*				
PHONEMIC > NON-PHONEMIC										
L	−60.5	−11.0	−3.6	−62.9	−12.9	−4.3	5.19	57	0.23	superior temporal gyrus
R	48.3	−29.9	−0.4	45.4	−37.7	5.4	4.02	33	0.15	superior temporal gyrus
R	59.5	−4.2	−6.8	58.3	−9.5	−9.2	4.81	34	0.18	superior temporal gyrus
NON-PHONEMIC > PHONEMIC										
R	48.4	−8.3	0.3	53.0	−1.9	4.7	4.88	69	0.16	planum polare (medial)
PHONEMIC > SINGLE-FORMANT										
L	−61.4	−17.9	−1.1	−62.9	−12.9	−4.3	7.27	190	0.27	ventral superior temporal gyrus
R	60.1	−0.9	−5.6	61.0	−2.7	−1.4	4.86	32	0.20	ventral superior temporal gyrus
SINGLE-FORMANT > PHONEMIC										
L	−42.6	−60.6	−6.4	−43.1	−66.7	2.5	4.85	42	0.08	inferior temporal sulcus
L	−39.7	−23.6	5.3	−40.9	−28.6	16.8	7.94	185	0.16	planum polare/temporale (medial)
R	43.4	−60.3	− − 1.1	45.7	−57.8	−0.0	4.45	47	0.10	inferior temporal sulcus
R	44.6	−19.4	7.6	56.0	−23.7	5.7	7.03	285	0.17	planum polare/temporale (medial)
SINGLE-FORMANT > NON-PHONEMIC										
L	−38.7	−32.9	13.8	−38.2	−32.2	11.3	6.49	59	0.11	planum temporale (medial)
STEADY-STATE ADAPTATION > BASELINE										
L	−62.3	−27.1	7.5	−65.1	−35.2	12.1	8.71	220	0.15	superior temporal gyrus (posterior)
L	−46.1	−2.6	−13.1	−41.4	−8.3	−11.5	5.54	60	0.15	superior temporal gyrus (anteior)
R	36.7	−82.1	11.1	38.0	−84.2	8.1	4.82	33	0.08	lateral occipital gyrus
R	55.5	−15.2	−3.1	52.8	2.1	−5.2	10.03	564	0.17	superior temporal gyrus
TRANSIENT ADAPTATION > BASELINE										
L	−59.6	−29.3	6.6	−59.9	−22.2	6.6	5.28	64	0.10	superior temporal gyrus
L	−58.5	−39.6	−9.2	−59.7	−43.7	−7.8	6.81	125	0.11	middle temporal gyrus
L	−41.8	−60.2	−10.5	−35.2	−65.3	−7.0	6.49	158	0.12	lateral occipital gyrus
L	−37.7	−72.4	13.0	−37.6	−70.6	12.4	5.49	43	0.09	inferior temporal sulcus
R	41.6	−66.7	4.4	38.0	−78.8	7.1	5.06	154	0.10	inferior temporal sulcus
R	54.9	2.1	−4.2	60.8	7.5	−6.3	5.47	37	0.14	superior temporal gyrus
STEADY-STATE AND TRANSIENT ADAPTATION > BASELINE										
L	−57.2	−24.2	6.7	−49.1	−27.9	4.5	7.71	286	0.14	superior temporal gyrus
R	58.8	−17.8	0.5	50.2	−2.8	−1.2	5.57	296	0.15	superior temporal gyrus
(STEADY-STATE ADAPTATION > BASELINE) X (SINGLE-FORMANT > PHONEMIC)										
R	41.2	−51.7	58.7	51.5	−52.0	53.9	28.79	36	0.56	inferior parietal sulcus
(STEADY-STATE ADAPTATION > BASELINE) X (SINGLE-FORMANT > NON-PHONEMIC)										
L	−57.0	−57.6	−21.8	−57.0	−56.7	−20.6	5.24	45	0.27	inferior temporal sulcus
(TRANSIENT ADAPTATION > BASELINE) X (PHONEMIC > SINGLE-FORMANT)										
L	−52.7	−2.6	−15.0	−49.6	−6.6	−17.9	6.18	47	0.09	superior temporal sulcus (anterior)
(STEADY-STATE AND TRANSIENT ADAPTATION > BASELINE) X (PHONEMIC > NON-PHONEMIC)										
R	56.7	−44.3	20.3	51.0	−44.3	15.7	5.68	56	0.08	superior temporal gyrus (posterior)
R	63.8	−49.5	5.4	64.3	−43.3	3.2	4.24	40	0.13	superior temporal sulcus (posterior)
(STEADY-STATE AND TRANSIENT ADAPTATION > BASELINE) X (SINGLE-FORMANT > NON-PHONEMIC)										
L	−59.1	−55.1	−2.3	−56.7	−62.0	−1.3	5.19	189	0.12	inferior temporal sulcus
L	−57.7	−29.1	13.5	−54.2	−34.1	17.9	5.75	41	0.08	superior temporal gyrus (posterior)
R	59.7	−34.4	6.6	56.2	−34.4	7.7	6.40	74	0.08	superior temporal gyrus (posterior)
R	60.4	−20.5	−11.4	63.9	−20.6	−10.3	5.55	33	0.12	middle temporal gyrus
TRANSIENT ADAPTATION > STEADY-STATE ADAPTATION										
L	−51.3	−50.7	38.1	−53.8	−53.2	36.8	5.55	31	0.11	super marginal gyrus
R	44.6	14.1	−37.2	47.0	13.9	−34.9	6.00	32	0.24	temporal pole
STEADY-STATE ADAPTATION > TRANSIENT ADAPTATION										
L	−45.4	−4.1	−6.7	−46.7	−6.0	3.4	7.15	40	0.15	superior temporal gyrus (anterior)
R	55.9	−22.3	0.8	53.2	−19.3	−1.2	9.72	52	0.12	superior temporal gyrus (posterior)

(Continued)

Table 1 | Continued

Hemi	Center			Peak			*t*-value	Cluster size (voxels)	Mean effect size (%)	Region
	X	*Y*	*Z*	*X*	*Y*	*Z*				
TRANSIENT ADAPTATION > STEADY-STATE ADAPTATION[a]										
L	−45.6	7.9	−33.8	−47.4	22.4	−26.4	4.60	113	0.16	temporal pole
L	−42.0	−66.7	27.7	−53.8	−53.2	36.8	5.55	380	0.10	lateral occipital gyrus
R	44.7	−69.0	13.8	40.8	−72.9	27.3	6.31	217	0.08	lateral occipital gyrus
STEADY-STATE ADAPTATION > TRANSIENT ADAPTATION[a]										
L	−62.0	−34.8	11.5	−70.6	−27.2	10.7	5.01	117	0.12	superior temporal gyrus (posterior)
L	−45.8	−0.8	−7.9	−46.7	−6.0	3.4	7.15	151	0.12	superior temporal gyrus (anterior)
R	47.6	2.9	−11.2	50.0	10.5	−3.7	5.63	111	0.13	superior temporal gyrus (anterior)
R	58.6	−25.3	2.9	53.2	−19.3	−1.2	9.72	209	0.10	superior temporal gyrus (posterior)

Unless noted, all contrasts are threshold = $p < 0.01$ (0.05 corrected).

[a] *threshold = $p < 0.05$ (0.05 corrected).*

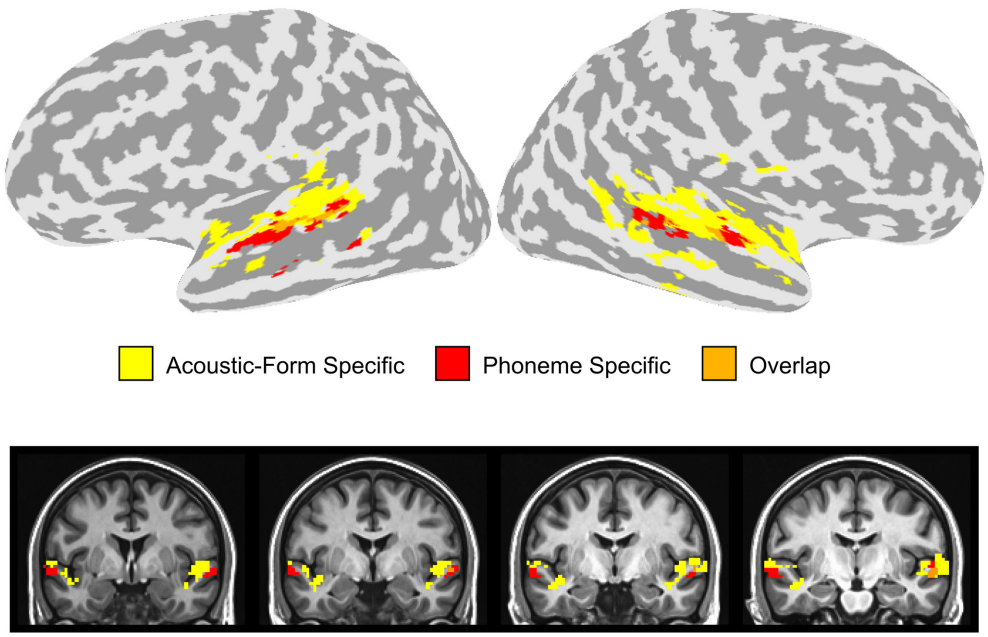


FIGURE 5 | Overlay of stimulus and adaptation effects in the STG. Voxels in the dorsal or ventral STG that were significantly active in the Phonemic > Single-Formant or Phonemic > Non-Phonemic contrasts (i.e., phoneme specific) are displayed in red. Voxels in the STG that were significant in any of the **three** adaptation contrasts (i.e., acoustic-form specific) are shown in yellow. Overlapping voxels are colored orange.

area for stimuli with more complex spectrotemporal structure, such as speech, animal vocalizations, or environmental sounds (Altmann et al., 2007; Joly et al., 2012; Lewis et al., 2012). In the current study, the bilateral dorsal STG demonstrated adaptation to the transition and steady-state portions of the stimulus regardless of whether the stimulus was phonemic or not, suggesting that it plays a role in representing certain types of spectrotemporal features that are relevant (but not exclusive) to phoneme perception, such as the multi-frequency harmonics that form the steady-state period or the rapid frequency sweeps that occur during the transition period of speech syllables.

Increased activity in the dorsal STG was observed for all three adaptation conditions compared to baseline, however, there were some differences in the patterns of activation. First, the activation clusters in the two conditions with a change in the steady-state period (SS and TSS) were larger than those for the transition condition (T). Second, direct contrasts between the T and SS conditions (**Figure 4**) showed greater activity for SS in bilateral anterior and posterior STG, suggesting that neurons encoding information about the steady-state period are located across the entire STG, while the transition period is primarily encoded by neurons in an areas confined to the middle STG lateral to HG. The steady-state and transition periods of the stimuli used in the experiment have different types of spectrotemporal structure. The transition period consists of relatively fast changes in spectral content, while the steady-state period has relatively little

spectral variation over time. It is possible that neural processing during these two time periods involves different populations of neurons, which are sensitive to different types of spectrotemporal features. Studies in monkeys suggest that neurons in more anterior cortical fields (R and AL) have longer latencies and longer sustained responses than the more centrally-located A1, suggesting that these neurons process acoustic information over longer time windows (Tian and Rauschecker, 2004; Bendor and Wang, 2008; Scott et al., 2011). If the anterior auditory neurons in human have similar windows of integration as in the monkey (>100 ms), then these neurons would be less sensitive to the fast temporal changes during the transition period, resulting in less adaptation in the T condition. It has been suggested that these anterior auditory fields form an auditory ventral stream, in which both acoustic and linguistic information is processed at increasing longer time scales (Rauschecker and Scott, 2009). In speech, much of the longer acoustic information (i.e., prosody) is derived by tracking pitch intonation, which is primarily determined from the vowel steady-state periods. Although these neurons might be less sensitive to fast temporal changes during the transition period, they might be optimally tuned to detecting changes in the steady-state period. In line with this view, is the finding that sentences with scrambled prosody show reduced activation compared to normally spoken sentences in bilateral anterior STG (Humphries et al., 2005). In addition to the anterior STG, the current study also found a similar activation pattern in the posterior STG. This set of areas is thought to be part of a dorsal auditory stream involved in sound localization and speech-motor coordination (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Liebenthal et al., 2013). Like the anterior areas, decreased sensitivity in the posterior STG to the transition period could be related to longer processing windows. In contrast, the finding of high activity levels for both the T and SS conditions in a section of the middle STG, adjacent to the ventral STG area that showed greater response to the Phonemic condition, suggests that these two types of acoustic features are important for phoneme processing.

Greater activation for the T condition was found in several areas outside of auditory cortex. It has been suggested that vowels and consonants contribute differently to speech perception, with vowels containing the majority of acoustic information about prosody and segmentation, and consonants providing linguistic-based information about lexical identity (Nespor et al., 2003). The activation differences between T and SS could also be related to this distinction. Greater sensitivity to the steady-state periods corresponding to vowels was found in purely auditory regions and greater sensitivity to the transition period corresponding to the consonant was found in parts of the cortex considered to be heteromodal and possibly involved lexical semantic processing.

Higher levels of activity in the bilateral ventral STG were seen for the Phonemic condition compared to the Non-Phonemic and Single-Formant sounds. This is consistent with findings from a large body of studies that have found greater activation in this area in response to speech syllables compared to non-speech auditory controls (Obleser et al., 2007; Leaver and Rauschecker, 2010; Liebenthal et al., 2010, 2005; Leech and Saygin, 2011; Woods et al., 2011). Furthermore, the left ventral STG has been shown to have categorical response to speech syllables varied along an acoustic continuum suggesting that this area is involved in abstract representations of sound (Liebenthal et al., 2005; Joanisse et al., 2007). In the current study, the Non-Phonemic and Single-Formant stimuli were synthesized with parameters very closely matching the spectrotemporal composition of the Phonemic stimuli. Thus, the observed differences in activation cannot be attributed simply to differences in acoustic form. The fact that this area did not respond to adaptation further supports the view that it encodes abstract representations of sound.

The results from the current study support the view that there are multiple hierarchical processing streams extending from primary auditory cortex to anterior, posterior, and lateral parts of the temporal lobe (Rauschecker et al., 1995; Kaas and Hackett, 2000; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). The dorsal and ventral parts of the STG observed in the current study represent two stages along these hierarchical pathways. Neurons in the dorsal STG encode information about complex spectrotemporal features by integrating across simpler acoustic features represented in earlier stages in the hierarchy in primary auditory cortex. The ventral STG, in turn, integrates information from the dorsal STG to build more complex representations related specifically to phonemic patterns. As the representations become more complex, they also become more abstract with reduced sensitivity to acoustic form, allowing categorical identification of acoustically varying sounds, such as speech phonemes. In addition to this dorsal/ventral hierarchy, the difference observed here between adaptation to the transition and steady-state segments of the stimuli suggests that there are important anterior-posterior differences in the superior temporal cortex beyond those associated with the dual-stream model of auditory processing. The results are consistent with the existence of several functional pathways tuned to different types of acoustic information, specifically only slow spectrally changing information in anterior and posterior STG and both slow and fast spectral information in the middle STG.

Finally, on the medial side of the STP, a larger response was found for Non-Phonemic and Single-Formant sounds compared to Phonemic sounds. This area did not activate in the adaptation contrasts. Other studies have observed a similar preference for non-speech over speech sounds in this region (Tremblay et al., 2013). Its location in medial auditory cortex suggests that it is homologous to the medial belt identified in the monkey. Interestingly, a study of the response properties of medial belt neurons in the monkey suggests a similar preference for spectral wide-band stimuli as in lateral belt neurons (Kuśmierek and Rauschecker, 2009). However, unlike lateral belt neurons, medial belt neurons do not show preferential responses to monkey vocalizations (Kuśmierek and Rauschecker, 2009). Thus, it is possible that the preference for non-phonemic sounds in medial auditory cortex could represent a tuning to sounds with unfamiliar, simpler harmonic structure.

In conclusion, we identified distinct regions of auditory cortex that were differentially sensitive to acoustic form and stimulus type, suggesting a hierarchical organization of auditory fields extending ventrolaterally from primary auditory cortex to the STS

and with varying sensitivity to acoustic form along the anterior to posterior axis of the STG. These results extend our understanding of the brain areas involved in auditory object identification and speech perception.

ACKNOWLEDGMENTS

This study was supported by funding from the National Institute on Deafness and Other Communication Disorders (R01 DC006287, Einat Liebenthal).

REFERENCES

Alho, K., Rinne, T., Herron, T. J., and Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. *Hear. Res.* 307, 29–41. doi: 10.1016/j.heares.2013.08.001

Altmann, C. F., Doehrmann, O., and Kaiser, J. (2007). Selectivity for animal vocalizations in the human auditory cortex. *Cereb. Cortex* 17, 2601–2608. doi: 10.1093/cercor/bhl167

Avants, B., and Gee, J. C. (2004). Geodesic estimation for large deformation anatomical shape averaging and interpolation. *Neuroimage* 23(Suppl. 1), S139–S150. doi: 10.1016/j.neuroimage.2004.07.010

Bendor, D., Osmanski, M. S., and Wang, X. (2012). Dual-pitch processing mechanisms in primate auditory cortex. *J. Neurosci.* 32, 16149–16161. doi: 10.1523/JNEUROSCI.2563-12.2012

Bendor, D., and Wang, X. (2008). Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *J. Neurophysiol.* 100, 888–906. doi: 10.1152/jn.00884.2007

Boemio, A., Fromm, S., Braun, A., and Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci.* 8, 389–395. doi: 10.1038/nn1409

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9, 179–194. doi: 10.1006/nimg.1998.0395

Grill-Spector, K., and Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol.* 107, 293–321. doi: 10.1016/S0001-6918(01)00019-1

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223.

Herdener, M., Esposito, F., Scheffler, K., Schneider, P., Logothetis, N. K., Uludag, K., et al. (2013). Spatial representations of temporal and spectral sound cues in human auditory cortex. *Cortex* 49, 2822–2833. doi: 10.1016/j.cortex.2013.04.003

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

Humphries, C., Love, T., Swinney, D., and Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Hum. Brain Mapp.* 26, 128–138. doi: 10.1002/hbm.20148

Joanisse, M. F., and Gati, J. S. (2003). Overlapping neural regions for processing rapid temporal cues in speech and nonspeech signals. *Neuroimage* 19, 64–79. doi: 10.1016/S1053-8119(03)00046-6

Joanisse, M. F., Zevin, J. D., and McCandliss, B. D. (2007). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. *Cereb. Cortex* 17, 2084–2093. doi: 10.1093/cercor/bhl124

Joly, O., Pallier, C., Ramus, F., Pressnitzer, D., Vanduffel, W., and Orban, G. A. (2012). Processing of vocalizations in humans and monkeys: a comparative fMRI study. *Neuroimage* 62, 1376–1389. doi: 10.1016/j.neuroimage.2012.05.070

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799. doi: 10.1073/pnas.97.22.11793

Kent, R. D. (2002). *Acoustic Analysis of Speech, 2 Edn.,* Clifton Park, NY: Cengage Learning.

Kuśmierek, P., and Rauschecker, J. P. (2009). Functional specialization of medial auditory belt cortex in the alert *Rhesus monkey*. *J. Neurophysiol.* 102, 1606–1622. doi: 10.1152/jn.00167.2009

Lawyer, L., and Corina, D. (2014). An investigation of place and voice features using fMRI-adaptation. *J. Neurolinguistics* 27, 18–30. doi: 10.1016/j.jneuroling.2013.07.001

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Leech, R., and Saygin, A. P. (2011). Distributed processing and cortical specialization for speech and environmental sounds in human temporal cortex. *Brain Lang.* 116, 83–90. doi: 10.1016/j.bandl.2010.11.001

Leff, A. P., Iverson, P., Schofield, T. M., Kilner, J. M., Crinion, J. T., Friston, K. J., et al. (2009). Vowel-specific mismatch responses in the anterior superior temporal gyrus: an fMRI study. *Cortex* 45, 517–526. doi: 10.1016/j.cortex.2007.10.008

Lewis, J. W., Talkington, W. J., Tallaksen, K. C., and Frum, C. A. (2012). Auditory object salience: human cortical processing of non-biological action sounds and their acoustic signal attributes. *Frontiers in Systems Neuroscience* 6:27. doi: 10.3389/fnsys.2012.00027

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., and Desai, A. (2013). Neural dynamics of phonological processing in the dorsal auditory stream. *J. Neurosci.* 33, 15414–15424. doi: 10.1523/JNEUROSCI.1511-13.2013

Moerel, M., Martino, F. D., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., et al. (2013). Processing of natural sounds: characterization of multipeak spectral tuning in human auditory cortex. *J. Neurosci.* 33, 11888–11898. doi: 10.1523/JNEUROSCI.5306-12.2013

Nelken, I., Fishbach, A., Las, L., Ulanovsky, N., and Farkas, D. (2003). Primary auditory cortex of cats: feature detection or something else? *Biol. Cybern.* 89, 397–406. doi: 10.1007/s00422-003-0445-3

Nespor, M., Peña, M., and Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue E Linguaggio*. 2, 203–229. doi: 10.1418/10879

Norman-Haignere, S., Kanwisher, N., and McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.* 33, 19451–19469. doi: 10.1523/JNEUROSCI.2880-13.2013

Obleser, J., Leaver, A. M., VanMeter, J., and Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Front. Psychol.* 1:232. doi: 10.3389/fpsyg.2010.00232

Obleser, J., Zimmermann, J., Meter, J. V., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cereb. Cortex* 17, 2251–2257. doi: 10.1093/cercor/bhl133

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

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

Rauschecker, J. P., and Tian, B. (2004). Processing of band-passed noise in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* 91, 2578–2589. doi: 10.1152/jn.00834.2003

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the *Macaque* nonprimary auditory cortex. *Science* 268, 111–114. doi: 10.1126/science.7701330

Recanzone, G. H. (2008). Representation of con-specific vocalizations in the core and belt areas of the auditory cortex in the alert *Macaque monkey*. *J. Neurosci.* 28, 13184–13193. doi: 10.1523/JNEUROSCI.3619-08.2008

Rimol, L. M., Specht, K., Weis, S., Savoy, R., and Hugdahl, K. (2005). Processing of sub-syllabic speech units in the posterior temporal lobe: an fMRI study. *Neuroimage* 26, 1059–1067. doi: 10.1016/j.neuroimage.2005.03.028

Saad, Z. S., Glen, D. R., Chen, G., Beauchamp, M. S., Desai, R., and Cox, R. W. (2009). A new method for improving functional-to-structural MRI alignment using local Pearson correlation. *Neuroimage* 44, 839–848. doi: 10.1016/j.neuroimage.2008.09.037

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* 10:e1003412. doi: 10.1371/journal.pcbi.1003412

Schönwiesner, M., Rübsamen, R., and Von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *Eur. J. Neurosci.* 22, 1521–1528. doi: 10.1111/j.1460-9568.2005.04315.x

Scott, B. H., Malone, B. J., and Semple, M. N. (2011). Transformation of temporal processing across auditory cortex of awake *Macaques. J. Neurophysiol.* 105, 712–730. doi: 10.1152/jn.01120.2009

Stephens, J. D. W., and Holt, L. L. (2011). A standard set of American-English voiced stop-consonant stimuli from morphed natural speech. *Speech Commun.* 53, 877–888. doi: 10.1016/j.specom.2011.02.007

Tian, B., and Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* 92, 2993–3013. doi: 10.1152/jn.00472.2003

Tremblay, P., Baroni, M., and Hasson, U. (2013). Processing of speech and non-speech sounds in the supratemporal plane: auditory input preference does not predict sensitivity to statistical structure. *Neuroimage* 66, 318–332. doi: 10.1016/j.neuroimage.2012.10.055

Wang, X., Lu, T., Bendor, D., and Bartlett, E. (2008). Neural coding of temporal information in auditory thalamus and cortex. *Neuroscience* 157, 484–494. doi: 10.1016/j.neuroscience.2008.07.050

Wolmetz, M., Poeppel, D., and Rapp, B. (2010). What does the right hemisphere know about phoneme categories? *J. Cogn. Neurosci.* 23, 552–569. doi: 10.1162/jocn.2010.21495

Woods, D. L., Herron, T. J., Cate, A. D., Kang, X., and Yund, E. W. (2011). Phonological processing in human auditory cortical fields. *Front. Hum. Neurosci.* 5:42. doi: 10.3389/fnhum.2011.00042

Zatorre, R. J., Belin, P., and Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends Cogn. Sci.* 6, 37–46. doi: 10.1016/S1364-6613(00)01816-7

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: 15 August 2014; accepted: 22 November 2014; published online: 11 December 2014.

Citation: Humphries C, Sabri M, Lewis K and Liebenthal E (2014) Hierarchical organization of speech perception in human auditory cortex. Front. Neurosci. ***8****:406. doi: 10.3389/fnins.2014.00406*

This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 September 2014
doi: 10.3389/fnins.2014.00289

The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults

Einat Liebenthal[1,2]*, Rutvik H. Desai[3], Colin Humphries[1], Merav Sabri[1] and Anjali Desai[1]

[1] *Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA*
[2] *Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, USA*
[3] *Department of Psychology, University of South Carolina, Columbia, SC, USA*

Edited by:
Josef P. Rauschecker, Georgetown University School of Medicine, USA

Reviewed by:
Mireille Besson, Institut de Neurosciences Cognitives de la Meditarranée, France
Peter E. Turkeltaub, Georgetown University, USA

****Correspondence:***
Einat Liebenthal, Functional Neuroimaging Laboratory, Brigham and Women's Hospital/Harvard Medical School, 824 Boylston Street, Chestnut Hill, MA 02467, USA
e-mail: einatl@mcw.edu; eliebenthal@partners.org

The superior temporal sulcus (STS) in the left hemisphere is functionally diverse, with sub-areas implicated in both linguistic and non-linguistic functions. However, the number and boundaries of distinct functional regions remain to be determined. Here, we present new evidence, from meta-analysis of a large number of positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) studies, of different functional specificity in the left STS supporting a division of its middle to terminal extent into at least three functional areas. The middle portion of the left STS stem (*f*mSTS) is highly specialized for speech perception and the processing of language material. The posterior portion of the left STS stem (*f*pSTS) is highly versatile and involved in multiple functions supporting semantic memory and associative thinking. The *f*pSTS responds to both language and non-language stimuli but the sensitivity to non-language material is greater. The horizontal portion of the left STS stem and terminal ascending branches (*f*tSTS) display intermediate functional specificity, with the anterior-dorsal ascending branch (*f*atSTS) supporting executive functions and motor planning and showing greater sensitivity to language material, and the horizontal stem and posterior-ventral ascending branch (*f*ptSTS) supporting primarily semantic processing and displaying greater sensitivity to non-language material. We suggest that the high functional specificity of the left *f*mSTS for speech is an important means by which the human brain achieves exquisite affinity and efficiency for native speech perception. In contrast, the extreme multi-functionality of the left *f*pSTS reflects the role of this area as a cortical hub for semantic processing and the extraction of meaning from multiple sources of information. Finally, in the left *f*tSTS, further functional differentiation between the dorsal and ventral aspect is warranted.

Keywords: functional organization, superior temporal sulcus (STS), left hemisphere, meta-analysis, functional magnetic resonance imaging (fMRI), positron emission tomography (PET), speech perception, semantic processing

INTRODUCTION

The human superior temporal sulci occupy an important fraction of the temporal cortex, strategically located at the junction of major temporal—parietal and—frontal functional pathways. Portions of the superior temporal sulcus (STS) in each hemisphere have been assigned numerous specialized perceptual and cognitive functions (Hein and Knight, 2008). Given the size and orientation of the STS, a division along its anterior-to-posterior axis is predicted, but determination of the functional boundaries remains hotly debated. Anatomically, the STS in each hemisphere has been divided into a forward stem composed of an anterior, a middle, a posterior and an horizontal segment, and a backward ascending branch bifurcated into an anterior-dorsal and a posterior-ventral segment, based on 3D morphology and ontogenic observations (Ochiai et al., 2004). In the left hemisphere, structural and functional connectivity patterns to the inferior frontal cortex support a division of the superior temporal cortex into at least two, and perhaps three, segments that are part of functionally distinct anterior-ventral and posterior-dorsal streams for language processing (Frey et al., 2008; Saur et al., 2008; Rauschecker and Scott, 2009; Rauschecker, 2011; Turken and Dronkers, 2011), reminiscent of the dual stream model of auditory perception (Rauschecker and Tian, 2000). Functional neuroimaging data also suggests that the left STS can be divided along its anterior-to-posterior axis, with the left middle STS consistently associated with speech perception (Liebenthal et al., 2005; Obleser et al., 2007; Leaver and Rauschecker, 2010; DeWitt and Rauschecker, 2012) and more posterior areas associated with multiple functions including semantic processing (Dronkers et al., 2004), audiovisual integration (Calvert et al., 2001; Beauchamp, 2005), biological motion processing (Puce et al., 1998) and phonological processing (Wise et al., 2001; Buchsbaum et al., 2005; Liebenthal et al., 2010, 2013). However, the different functions associated with different portion of the left STS have seldom been localized and compared within the same set of subjects and experimental framework. Previous studies of the STS have compared pairs of similar functions within a cognitive domain, such as for example voice and speech

recognition (Belin et al., 2000), speech perception and phonological processing (Liebenthal et al., 2010), or auditory, visual and somatosensory integration (Beauchamp et al., 2008). But, to our knowledge, systematic functional comparisons have not been carried out between multiple functions across cognitive domains (for example, between several language and non-language functions). As a result, the number and boundaries of distinct functional regions in the left STS remain to be determined.

Despite a remarkable growth in neuroimaging research in recent years, another persistent limitation to understanding the neuroanatomical organization of cognitive functions is that most studies rely on relatively small sample sizes and narrow experimental designs (i.e., a restricted number of experimental conditions). This is problematic because of the well-known inter-individual variability in brain structure, brain function, and brain structure-function relationships, including in the STS (Sowell et al., 2002; Kanai et al., 2012; Gilaie-Dotan et al., 2013). Particularly in the terminal aspect of the STS, the number of ascending branches and how they join the STS stem was found to be highly variable between individuals, causing irregularity in naming convention and contributing to the murkiness in functional characterization of this area (Segal and Petrides, 2012). Further challenging the characterization of terminal STS is the high degree of variability in the neighboring inferior parietal lobule (IPL), where the supramarginal gyrus (SMG) and angular gyrus (AG) were found to be composed of several distinct cytoarchitectural areas, suggestive of functional differentiation, with no consistent correspondence between cytoarchitectural and macroanatomical borders (Caspers et al., 2006). It is therefore valuable to examine brain activation patterns *across* neuroimaging studies in order to identify reliable functional organization principles in larger subject samples and in a wide array of cognitive paradigms.

Previous meta-analyses involving the temporal cortex have most often centered on one specific cognitive function, for example speech perception (Turkeltaub and Coslett, 2010), semantic processing (Binder et al., 2009; Adank, 2012), auditory attention (Alho et al., 2014), writing (Purcell et al., 2011; Planton et al., 2013), motion perception (Grosbras et al., 2012), emotion processing (Lee and Siegle, 2012), and theory of mind (Van Overwalle and Baetens, 2009). One prior meta-analysis focused on the multi functionality of the STS, but was limited to just a few studies per functional category that used similar stimuli and experimental designs (Hein and Knight, 2008).

The present meta-analysis was designed to study the functional organization of the left STS for language and non-language processing. The meta-analysis deliberately included a large number of studies using different neuroimaging methods (PET, fMRI), experimental designs (implicit, explicit, or no task), and stimuli (linguistic, nonlinguistic). The extent of the left STS was determined based on a probabilistic map created from structural magnetic resonance (MR) images of 61 brains. We reasoned that (1) drawing from commonalities in activation across multiple data sets generated using different experimental designs and methodologies would highlight reliable and fundamental functional organization patterns; and (2) defining the extent of the STS and a comprehensive set of putative STS functional categories would serve as a unifying platform for analyzing results from multiple studies, irrespective of anatomical labeling practices and interpretation of functional activation patterns across the studies. The reported results rely on analysis of 485 activation peaks from 253 studies that fell within the left STS mask. The peaks were sorted into 2 stimulus categories and 15 functional categories based on the experimental contrast used to generate each activation map. The main results are reported in terms of functional specificity, expressed as the number of stimulus and functional categories with a significant mean activation likelihood estimate, in different areas of the left STS. Structural subdivisions of the STS are labeled using an approximation of the demarcation of Ochiai et al. (2004), as detailed schematically in **Figure 1A**. Note that the anterior-dorsal ascending branch of the terminal STS (atSTS) is immediately posterior to the ascending branch of the Sylvian fissure. The atSTS is expected in most brains to be anterior to the first intermediate sulcus of Jensen, which (when present) is considered to form the boundary

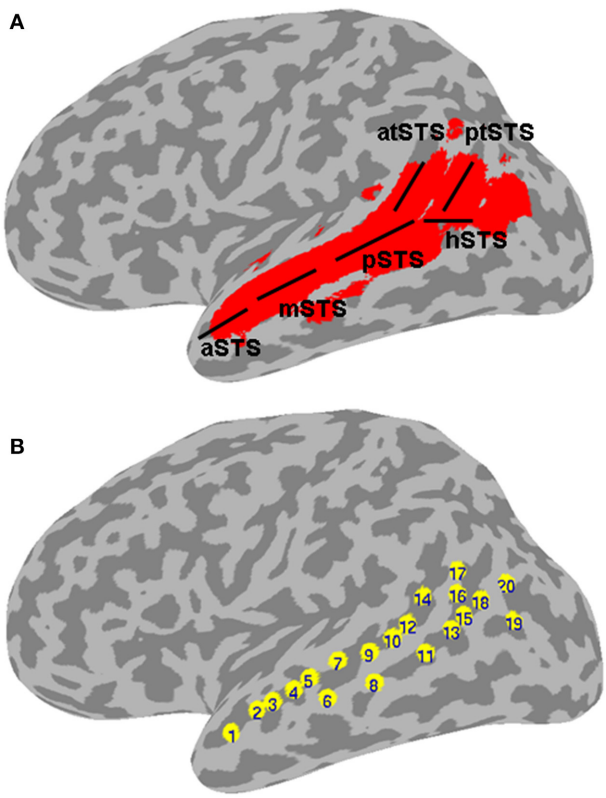


FIGURE 1 | Left STS probabilistic mask and ROIs. (A) Probabilistic mask of the left STS (in red) shown projected onto a cortical surface model of the Colin brain in Talairach space. Also shown is a schematic approximation of the STS anatomical subdivisions used to describe the results (based on Ochiai et al., 2004), consisting of the aSTS, anterior STS; mSTS, middle STS; pSTS, posterior STS; hSTS, horizontal STS; atSTS, anterior branch of terminal STS; and ptSTS, posterior branch of terminal STS; **(B)** twenty evenly spaced spherical ROIs in the left STS, in which functional specificity was probed, shown projected onto the same cortical surface. The ROIs are numbered in ascending order according to their anterior-posterior position along the left STS.

between the SMG and AG (Caspers et al., 2006; Segal and Petrides, 2012). As such, the atSTS terminates in most brains within the posterior SMG, near the boundary with AG. The posterior-ventral branch of the terminal STS (ptSTS) terminates within the AG.

MATERIALS AND METHODS

A probabilistic map of the left STS was created by averaging two T1-weighted MR images from each of 61 brains, in which the STS had been demarcated using Freesurfer (Dale et al., 1999) for automatic parcellation of sulci and gyri (Destrieux et al., 2010). The resulting STS atlas (labeled TT_desai_ddpmaps) is included with AFNI (Cox, 1996). The left STS probabilistic map was thresholded at 20% probability and extended 5 mm laterally to create a mask for the meta-analysis (**Figure 1A**). Note that the STS, as parcellated in the Destrieux et al. atlas, broadens toward the posterior end and arguably includes parts of the posterior middle temporal gyrus (pMTG), AG, and possibly SMG. We chose to use the same parcellation for consistency and to ensure adequate sampling of activation in the terminal STS.

In the BrainMap database (Laird et al., 2005), 675 PET and fMRI studies published in the years 1990–2010 were identified that reported activation peaks located within the left STS mask, as assessed based on reported coordinates in Talairach space (Talairach and Tournoux, 1988). From these, 485 activation peaks from 253 different studies meeting the inclusion criteria of representing data collected from a group of at least 8 healthy adults of mixed gender, and using a high-level baseline, were incorporated in the meta-analysis. Functional contrasts using a low-level baseline, such as fixation or rest, were excluded due to the uncertainty associated with the nature of activations in such comparisons.

Each activation peak was categorized according to the type of stimulus material and the functional contrast used to generate the activation. The stimulus categories consisted of "language" (including auditory and visual spoken, or written, sub-syllabic, syllabic, word, sentence or discourse stimuli) and "non-language" (including all types of non-verbal and non-written stimuli not included in the language category). The functional categories consisted of 15 sensory, motor, or cognitive processes most commonly targeted by the condition contrasts used to generate the peaks included in the meta-analysis. The functional categories were further classified as linguistic or non-linguistic for the purpose of comparing each functional category with the other categories in its class. The complete list of stimulus and functional categories, and functional classes, is given in **Table 1**.

Peaks were assigned to a stimulus category based on the input material used in the "high" (of interest) compared to "low" (baseline) condition of the experimental contrast, and to up to three different functional categories representing the main sensory, motor, or cognitive functions considered to be engaged in the "high" relative to "low" condition of the contrast. For example, an activation peak resulting from a perceptual contrast of clear spoken sentences and non-intelligible speech-like sounds would be assigned to the language stimulus category and to the functional categories of speech perception, semantic processing, and syntactic processing. There were 271 and 223 peaks assigned to the language and non-language stimulus categories, respectively. Sixteen peaks were assigned to both the Language and Non-Language stimulus categories. These peaks resulted from contrasts in which the stimuli used in the "high" condition contained both linguistic and non-linguistic information that was not balanced by the stimuli used in the "low" condition. For example, some studies of audiovisual speech perception compared a video clip of a face producing natural speech with a series of stilled frames of the face showing apical gestures (Calvert and Campbell, 2003). The differential activation in this contrast was considered to reflect the higher linguistic (speech) and non-linguistic (biological motion) content of the stimuli in the "high" condition. Seven peaks were not assigned to either Language or Non-language stimulus categories. These peaks resulted from contrasts in which no stimulus was used in the "high" condition. For example, some studies compared an internal task such as imagination, in which no external stimulus was used, with a perceptual task (Kosslyn et al., 1996). Such peaks were assigned to functional categories and were included in comparisons between functional (but not stimulus) categories. The number of peaks assigned to each functional category (reported in **Table 1**) ranged 14–118 (mean = 37), with "semantic processing" as the largest category. The degree of overlap in peak assignment between pairs of functional categories ranged 0.03–0.52 (mean = 0.23), with the largest overlap occurring between "orthographic processing" and "semantic processing."

The GingerALE version 2.0.4 application of the BrainMap software was used to perform the meta-analysis, with fixed 10 mm FWHM Gaussian smoothing (Turkeltaub et al., 2002; Eickhoff et al., 2009, 2012). The activation likelihood estimation (ALE)

Table 1 | Stimulus and functional categories used to sort the left STS activation peaks.

Stimulus Categories	Number of peaks analyzed	Functional Categories	Number of peaks analyzed	Functional Class
Language	271	Orthographic processing	25	Language
		Phonological processing	20	
		Semantic processing	118	
		Speech perception	42	
		Speech production	17	
		Syntactic processing	21	
Non-language	223	Attention	25	Non-language
		Auditory processing	14	
		Biological motion processing	14	
		Emotion processing	87	
		Executive control	32	
		Memory	72	
		Motor control/planning	20	
		Multisensory processing	19	
		Visual processing	32	

*Each activation peak was categorized according to the type of stimulus material (language or non-language, in pink or orange shading, respectively) and function (sensory, motor, and cognitive) engaged in the "high" (of interest) relative to "low" (baseline) condition of the experimental contrast. The functional categories were further classified as linguistic (gray shading) or non-linguistic (blue shading) for the purpose of comparing each functional category with the other categories in its class (in **Figures 2**, **3**). The number of peaks analyzed in each stimulus and functional category in also reported.*

technique estimates the convergence of neuroimaging activation foci by modeling them as Gaussian probability distributions based on assessment of spatial uncertainty due to intersubject and co-registration variability. A relatively low and fixed (i.e., not adjusted according to study sample size) level of smoothing was used in order to maintain sensitivity to potential small subdivisions within the STS and to avoid potential bias from systematic differences in study sample sizes across functional categories. The ALE in the two stimulus categories was compared (**Figure 2A**). The ALE in each functional category was compared with the ALE in *all* other functional categories in the same class (**Figure 2B**), and also with the ALE in *each* of the other functional categories in the same class in a pairwise fashion (**Figure 3**), where class was defined as language or non-language (see **Table 1**). The ALE category contrast maps for the entire left STS were thresholded at $p < 0.01$ and clusters smaller than 700 μl were removed, resulting in a corrected error probability of $\alpha < 0.05$, as determined using the AlphaSim module in AFNI (Ward, 2000).

In a second analysis, the functional organization of the left STS was studied in finer grain by using a region of interest (ROI) approach. The left STS mask was divided into twenty ROIs. Because the geometry of the STS does not follow a straight line, we used a clustering algorithm to partition the left STS mask into twenty sub-regions that were approximately equal-sized and evenly spaced. This was accomplished by submitting the x, y, z coordinates of all the voxels in the mask to a k-means clustering algorithm set to identify twenty clusters. The cluster center coordinates were then used as the center positions of twenty 4 mm-radius spherical ROIs. The location of ROIs within the left STS mask is shown in **Figure 1B**. The mean ALE (expressed in *z*-scores) within each ROI was calculated for each functional category. The functional specificity of each ROI was estimated by tallying the number of categories activating this region at $p < 0.005$ ($z > 2.807$). Results of the ROI analyses are shown in **Figures 4**, **5**.

The cortical inflated surfaces in **Figure 2** were rendered using Caret 5.62 (Van Essen et al., 2001). The surfaces in the other figures were rendered using custom code in Matlab (Matlab 7.1, The Math Works Inc., Natick, MA).

RESULTS

The contrast between the two stimulus categories (**Figure 2A**) showed a greater likelihood of language compared to non-language activation peaks in most of the left STS, except in the posterior and horizontal STS stem (pSTS and hSTS, respectively) where a greater likelihood of non-language activation peaks was observed. The contrasts between each functional category in the language class and all the other categories in that class (**Figure 2B**, left panels) revealed significantly greater likelihood of speech perception peaks in the middle STS stem (mSTS), and of semantic processing peaks in pSTS and hSTS. The contrasts between each functional category in the non-language class and all the other categories in that class (**Figure 2B**, right panels) revealed significantly greater likelihood of emotion processing peaks in pSTS, and of executive processing peaks in the anterior terminal STS branch (atSTS). The non-language area in the stimulus contrast (**Figure 2A**) overlapped considerably with the semantic and emotion areas in the functional contrasts (**Figure 2B**). Pairwise comparisons between the functional categories in each class (**Figure 3**) revealed greater likelihood of speech perception peaks in mSTS relative to greater likelihood of speech production peaks in the anterior (atSTS) and posterior (ptSTS) terminal STS branches, as well as greater likelihood of auditory perception peaks in mSTS relative to greater likelihood of visual perception peaks in pSTS and hSTS.

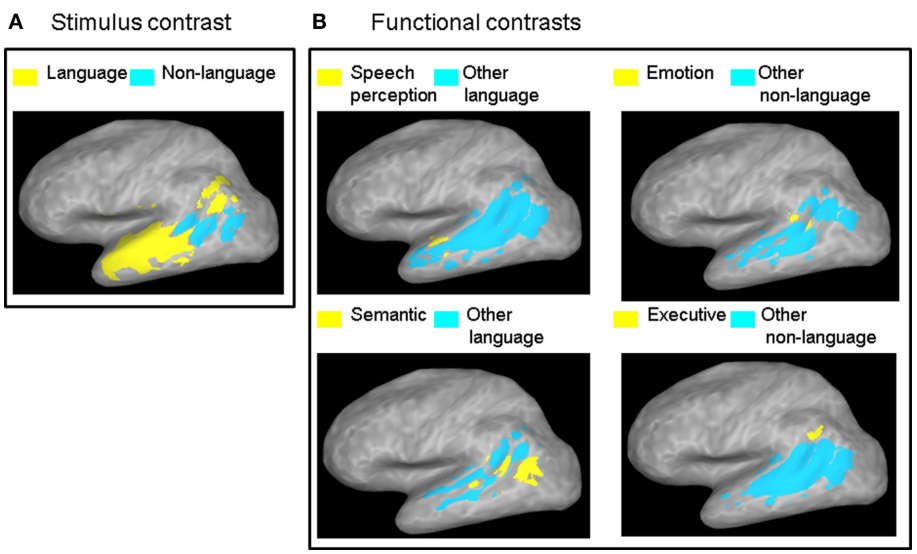


FIGURE 2 | ALE contrast maps. Contrast maps of **(A)** the two stimulus categories and **(B)** each functional category relative to all the other functional categories in the same class. Maps are thresholded at a corrected probability of $\alpha < 0.05$. Only functional contrasts resulting in significant differential ALE measurement in the left STS are displayed.

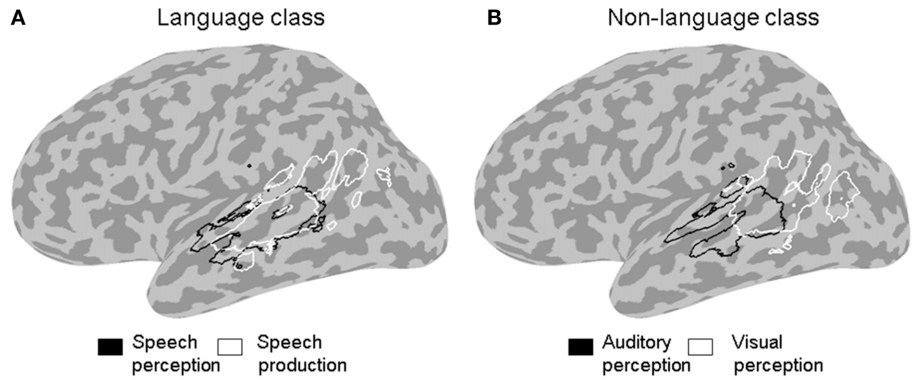


FIGURE 3 | ALE outline maps. Outline maps of pairs of functional categories in the language **(A)** and the non-language **(B)** functional classes that showed significant differences in activation likelihood in the left STS at a corrected error probability of $\alpha < 0.05$.

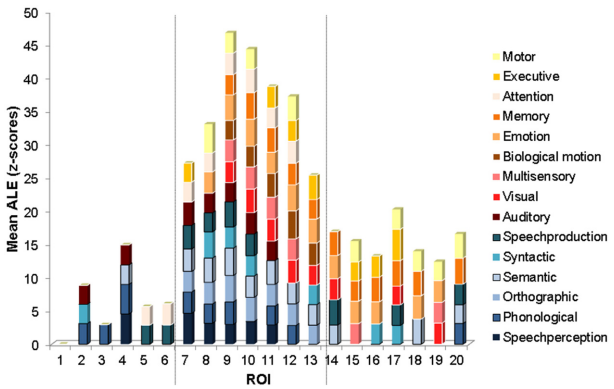


FIGURE 4 | ROI mean ALE functional measure. Graph of mean ALE measure within each ROI for each functional category exceeding the significance level (mean $z = 2.807$, $p < 0.005$). The categories are stacked according to functional class, with the language functional categories on the bottom in shades of blue, and the non-language functional categories on the top in shades of red and orange. The ROIs are ordered from the most anterior (ROI 1) to the most posterior (ROI 20) along the STS (see **Figure 1B** for the anatomical location of each ROI). The vertical dashed lines show locations of marked changes in functional specificity.

Other functional contrasts resulted in no significant differences ($\alpha < 0.05$).

The ROI analysis revealed the functional properties of the left STS with greater spatial detail. The mean ALE within each ROI, for each of the functional categories is plotted in **Figure 4**. Several interesting observations arise from this analysis. The most anterior ROIs (numbered 2–6) show significant activation likelihood for just a few functional categories (range 1–4, mean 2.6) largely from the language class (in shades of blue). The ROIs in intermediate position (numbered 7–13) show significant activation likelihood for the largest number of functional categories (range 8–14, mean 11) from both the language and non-language classes (the latter in shades of red and orange). The ROIs in the most posterior part of the left STS (numbered 14–20) show significant activation likelihood for an intermediate number of functional categories (range 4–6, mean 4.7) from both the language and non-language classes. The difference in functional specificity (expressed as the number of functional categories with a significant mean ALE measure) between the three regions is significant [One-Way ANOVA, $F_{(2,\ 16)} = 43$, $p = 0$]. Anatomically, the anterior ROIs (2–6) correspond roughly to the mSTS stem area, the intermediate ROIs (7–13) correspond roughly to the pSTS stem area, and the most posterior ROIs correspond roughly to the hSTS stem area and the atSTS and ptSTS branches. Note that in ROI 1, none of the categories survived the statistical threshold, likely due to a small number of activation peaks falling within this area.

Based on these differences in functional specificity, we propose a division of the left STS into middle, posterior and terminal functional areas, labeled respectively *f*mSTS (talairach y coordinates −7 to −27), *f*pSTS (talairach y coordinates −28 to −59), and *f*tSTS (talairach y coordinates −55 to −71). **Figure 5** shows an approximate demarcation of the three functional areas and their specificity, as well as plots of the mean ALE measure for each stimulus and functional category in the ROIs activated by the largest number of functional categories (i.e., the least specific ROIs) in each sub-division. In the *f*mSTS, the least functionally specific ROI (number 4) showed significant activation likelihood only for language stimuli, and only for the speech perception, and phonological, auditory, and semantic processing functional categories. In the *f*pSTS, the least specific ROI (number 9) showed significant activation likelihood for both language and non-language stimuli, and for 14 out of the 15 possible functional categories (with the exception of executive control). In the *f*tSTS, the least functionally specific ROI (number 17) showed significant activation likelihood for both language and non-language stimuli, and for the executive and motor control, memory, speech production, and syntactic and visual processing functional categories.

With regard to *f*tSTS, despite the similar level of functional specificity across this area, we expect that it is composed of an anterior and a posterior subdivision (*f*atSTS and *f*ptSTS, respectively), based on its irregular 3D anatomy and apparent dichotomous functionality related primarily to executive control in atSTS and to semantic processing in hSTS and ptSTS (see **Figure 2**).

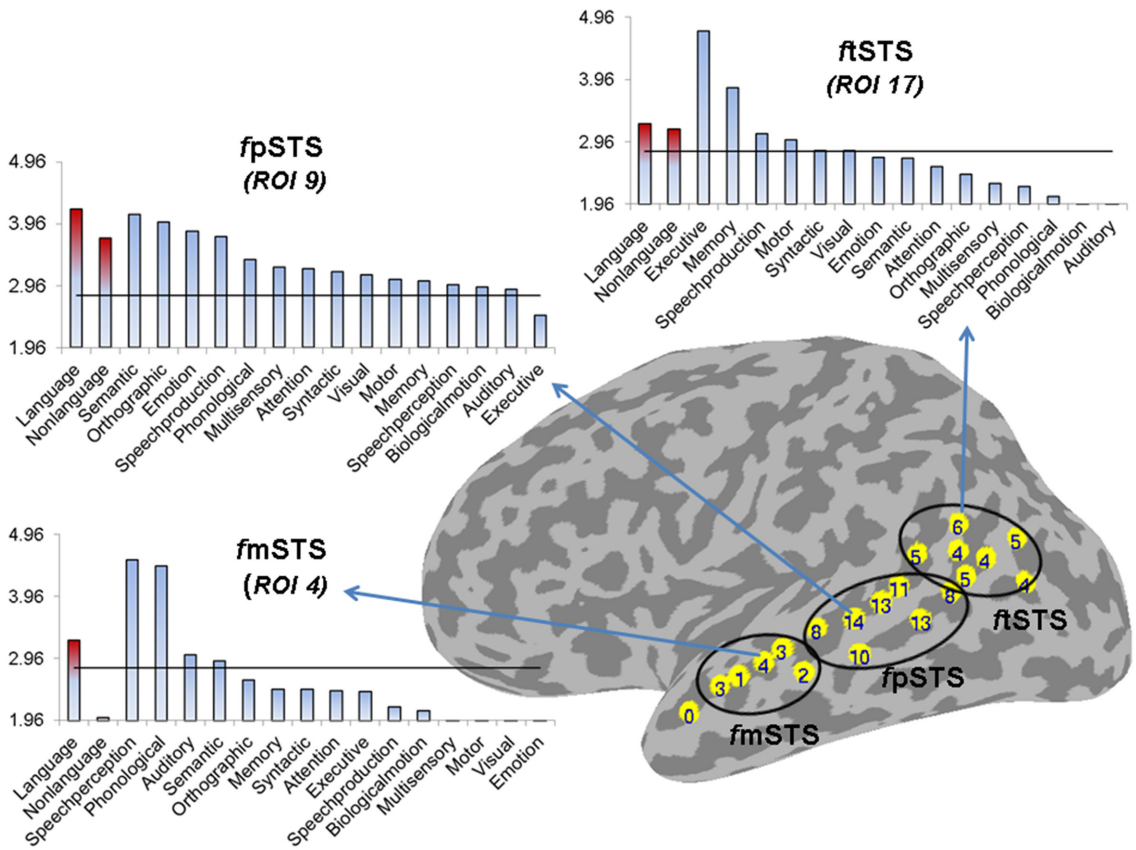


FIGURE 5 | Partition of left STS into three subdivisions based on functional specificity. The number label within each ROI represents its functional specificity, expressed as the number of functional categories with a significant mean ALE measure in this region ($p < 0.005$). The functional mSTS (*f*mSTS) subdivision was defined as a region activated by a small number of functional categories (range 1–4, mean 2.6), the functional pSTS (*f*pSTS) subdivision was defined as a region activated by the largest number of functional categories (range 8–14, mean 11), and the functional tSTS (*f*tSTS) subdivision was defined as a region activated by an intermediate number of functional categories (range 4–6, mean 4.7). The three graphs show the mean ALE measure (expressed in Z-scores) for each stimulus (in red) and functional (in blue) category in descending order of magnitude, in the ROIs that were activated by the largest number of functional categories in each subdivision (ROIs number 4, 9, and 17 in the left *f*mSTS, *f*pSTS, and *f*tSTS, respectively). The horizontal line corresponds to $z = 2.807$ ($p < 0.005$).

Several potential limitations should be mentioned with respect to the results. First, the Brainmap database is not a random sample of the neuroimaging literature and may be biased toward studies of certain cognitive functions. For example, the smaller number of studies of speech perception (42) compared to studies of semantic processing (118) found here with peaks falling in the left STS may reflect a sampling bias in the database or a true aspect of STS functional organization. Seconds, the distribution of number of peaks analyzed was not even along the left STS, with fewer peaks falling in the mSTS area (66) and more peaks falling in the pSTS (224) and tSTS (195) areas. Importantly, the difference in the distribution of the number of peaks along the STS cannot in itself explain the higher functional specificity of the mSTS because the distribution of the number of peaks was not random with respect to functional and stimulus category. That is, not all the stimulus and functional categories were evenly less represented in mSTS relative to pSTS and tSTS. On the contrary, a small number of categories were actually better represented in mSTS than in the rest of the STS. In particular, the category of speech perception had higher ALE values than all of the other language categories combined specifically in mSTS (**Figure 2B**), and the mSTS showed higher ALE values for Language over Non-Language stimuli (**Figure 2A**).

DISCUSSION

We present here new evidence from meta-analysis of a large number of PET and fMRI studies, of different functional specificity along the left STS supporting a division of its middle to terminal extent into at least three functionally distinct areas. Based on the present results, and a review of the literature, we suggest that a functional area in the left middle STS (*f*mSTS; Talairach y coordinates −7 to −27) is highly specialized for speech perception and the processing of language material. A functional area in the left posterior STS (*f*pSTS; Talairach y coordinates −28 to −59) is highly versatile and serves as a hub for semantic processing and multiple functions supporting semantic memory and associative thinking. The *f*pSTS responds to both language and non-language stimuli but the likelihood of response to non-language

material is greater. A functional area including the left horizontal and terminal STS (*f*tSTS; Talairach y coordinates −55 to −71) displays intermediate functional specificity, with the anterior ascending branch adjoining SMG (*f*atSTS) supporting executive functions and motor planning and showing greater likelihood of response to language material, and the horizontal stem and posterior ascending branch adjoining AG (*f*ptSTS) supporting primarily semantic processing and displaying greater likelihood of response to non-language material. These latter results in the *f*tSTS suggest that a further functional differentiation between its dorsal and ventral aspect is warranted.

The finding of a strong convergence of activity related to speech processing in the left *f*mSTS is largely consistent with prior neural functional models associating this area with phonemic perception (Davis and Johnsrude, 2003; Liebenthal et al., 2005; Obleser et al., 2007; Leaver and Rauschecker, 2010; DeWitt and Rauschecker, 2012). The left mSTS is considered to be part of a ventral auditory pathway for speech recognition, connecting the auditory cortex to semantic regions widely distributed in the left middle and inferior temporal cortex. Neurons in the left mSTS may be specially tuned to the categorical properties of native speech phonemes (Liebenthal et al., 2005; Leaver and Rauschecker, 2010; Humphries et al., 2013) making this area critical for decoding incoming speech signals. The most novel and striking aspect of the current results is the narrow functional specificity of the left *f*mSTS, observed as significant preference to language over non-language stimuli and to speech perception over other language functions (**Figures 2**, **3**), as well as the convergence of peaks from only a few functional categories mostly in the language class (**Figures 4**, **5**), in this area. It is possible that the high functional specificity of the left *f*mSTS for speech is an important means by which the human brain achieves its exquisite affinity and efficiency for native speech perception. The anatomical proximity of the mSTS to auditory cortex, and higher sensitivity of this region to auditory over visual processing (**Figure 3**), are also consistent with a specialization in this area for speech perception over other (non-auditory based) language functions.

The finding of a strong convergence of activity related to semantic processing in the left *f*pSTS is consistent with prior work indicating the importance of the adjacent left posterior MTG (pMTG) to language comprehension (Price, 2000, 2010; Dronkers et al., 2004; Binder et al., 2009). Lesions in the left pMTG are known to be particularly detrimental to language comprehension (Boatman et al., 2000; Dronkers et al., 2004; Baldo et al., 2013). The left posterior superior temporal cortex is activated during language comprehension irrespective of the input modality, including during sign language processing in native signers (Bavelier et al., 1998; MacSweeney et al., 2006). The main novel aspect of the present results is again related to functional specificity, which was astonishingly low in the left *f*pSTS and in sharp contrast to the high functional specificity observed in the left *f*mSTS. The left *f*pSTS was found to be extremely multi-functional, being more likely to respond to non-language stimuli, during semantic and emotion processing over other language and non-language functions, respectively (**Figure 2**); but also likely to respond to language stimuli and to almost all other functional categories (**Figures 4**, **5**). The observation that an area "specializing" in semantic processing is overall more responsive to non-linguistic (i.e., non-verbal and non-written) stimuli is perhaps not intuitive. However, this finding is consistent with the idea that the very nature of semantic processing involves association of input from the different senses, analyzed in various ways (e.g., sensory features, biological motion, emotional valence, etc. . .), to extract information relevant to object recognition and comprehension. The extreme multi-functionality of the left *f*pSTS may reflect the role of this area as a cortical hub for semantic processing and the extraction of meaning from multiple sources of information. The strategic location of the left *f*pSTS, at the confluence of auditory and visual afferent streams, and fronto-parietal somato-motor and executive control efferent streams, is ideal for a cortical hub, in line with the concept of a neural convergence zone (Damasio, 1989; Meyer and Damasio, 2009) or epicenter (Mesulam, 1990, 1998).

The finding of a mixed pattern of functionality in the left *f*tSTS is perhaps not surprising given the complex anatomy of this area and varied functionality of bordering areas. The atSTS branch terminates near the SMG, an area suggested to serve as an auditory-motor interface (Guenther et al., 2006; Hickok and Poeppel, 2007), whereas the ptSTS branch terminates into the AG, an area associated primarily with semantic processing (Binder et al., 2009; Price, 2010). The preference observed here of the left *f*atSTS for language stimuli and executive and motor control functions (**Figures 2**, **3**) is well in line with the implication of this and the neighboring SMG area in phonological processing (Paulesu et al., 1993; Caplan et al., 1997; Wise et al., 2001; Buchsbaum et al., 2005; Buchsbaum and D'Esposito, 2009; Liebenthal et al., 2013) and the learning of ambiguous or non-native sound categories (Callan et al., 2004; Golestani and Zatorre, 2004; Raizada and Poldrack, 2007; Desai et al., 2008; Liebenthal et al., 2010; Kilian-Hutten et al., 2011). The *f*atSTS may be important for maintenance of auditory sequences in short-term memory while their auditory, somatosensory, and motor properties are analyzed to support phonemic perception. In contrast, the preference observed here of the left *f*ptSTS for non-language stimuli and semantic processing bears resemblance to the preference of the nearby *f*pSTS area, and is well in line with the implication of the AG in semantic retrieval and semantic integration (Price, 2000, 2010; Dronkers et al., 2004; Binder et al., 2009; Binder and Desai, 2011; Bonner et al., 2013). The left *f*ptSTS area could be an extension of the left *f*pSTS semantic area identified here and a functional bridge to the AG. Taken together, these results support a functional differentiation between the anterior-dorsal and posterior-ventral aspects of tSTS, in line with the different role of dorsal and ventral portions of the IPL. Nevertheless, given the documented high intersubject variability in terminal STS, caution should be used in treating differences in activation within this area and with the adjacent IPL. The functional differentiation within terminal STS should be addressed further in future work, perhaps taking into account cyoarchitectural information.

Structural connectivity and resting state functional connectivity patterns in the left temporal cortex are also in line with a left STS anterior-to-posterior segregation based on functional specificity. Disparate language pathways are thought to connect the left

middle and posterior superior temporal cortex with the inferior frontal gyrus (IFG), consistent with ventral and dorsal streams of processing for language (Saur et al., 2008; Rauschecker and Scott, 2009; Rauschecker, 2011). Structural connectivity measured with diffusion tensor imaging showed that the middle superior temporal cortex is connected to the anterior IFG via the ventral portion of the extreme capsule fiber system and also via the uncinate fasciculus. In contrast, the posterior superior temporal cortex is connected to the posterior IFG directly via the arcuate fasciculus, and also indirectly through the inferior parietal cortex via the superior longitudinal fasciculus (Catani et al., 2005; Parker et al., 2005; Anwander et al., 2007; Frey et al., 2008). The left pMTG was found to have particularly rich structural connections with other brains areas through several major pathways connecting it to the AG and to the rest of the temporal cortex, in addition to IFG (Turken and Dronkers, 2011). Similarly, resting state functional connectivity in the left middle superior temporal cortex was found to be limited to the posterior temporal cortex and the IFG (Turken and Dronkers, 2011). In contrast, functional connectivity in the left pMTG was found to be among the highest in the cerebral cortex, with connections to the left AG, anterior STG, and IFG (Buckner et al., 2009; Turken and Dronkers, 2011). The locus of most extensive functional connectivity in the left pMTG indicated in the Buckner study (Talairach x, y, z coordinates −62, −38, −12) coincides with the anterior-posterior position of the pSTS area of least functional specificity observed in the present study (ROI 9, Talairach x, y, z coordinates −48, −39, −1).

The current STS meta-analysis extends that of Hein and Knight (2008) by introducing a new functional specificity measure highlighting the organization of the left STS for language and non-language processing. This new perspective was possible mainly thanks to the much larger number of studies across language and non-language domains analyzed here. In the Hein and Knight study, activation peaks in the speech perception category were clustered in the anterior portion of the STS (approximately corresponding to the mSTS area described here), whereas those for several other categories (multisensory processing, biological motion processing) were clustered in the posterior portion the STS (approximately corresponding to the pSTS and tSTS areas described here) though with a small presence also in the anterior STS. The results were interpreted as different degrees of multi-functionality in the anterior and posterior STS rather than a functional differentiation *per se*, because there was some degree of spatial overlap between functional categories along the entire STS. The present meta-analysis supports the concept of differences in multi-functionality along the STS. But the extreme low multi-functionality in the mSTS and contrastingly extreme high multi-functionality in the adjacent pSTS observed here suggest that there may be fundamental differences between these areas reflecting a true functional specialization for speech perception and semantic processing, respectively, rather than merely a gradient of multi-functionality.

In conclusion, the present work demonstrated a division of the mid-to-terminal left STS into at least three functional areas based on functional specificity. Future work using a more detailed definition of stimulus and functional categories, as well as finer anatomic parcellation of the STS mask, may yield further insights into the functional organization of left STS and the interaction of each functional subdivision with neighboring regions. A comparison with the functional organization of the right STS is also warranted.

ACKNOWLEDGMENTS

The work was supported by NIH/NIDCD R01 DC006287 (Einat Liebenthal) and NIH R01 DC10783 (Rutvik H. Desai).

REFERENCES

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: two activation likelihood estimation (ALE) meta-analyses. *Brain Lang.* 122, 42–54. doi: 10.1016/j.bandl.2012.04.014

Alho, K., Rinne, T., Herron, T. J., and Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. *Hear. Res.* 307, 29–41. doi: 10.1016/j.heares.2013.08.001

Anwander, A., Tittgemeyer, M., von Cramon, D. Y., Friederici, A. D., and Knosche, T. R. (2007). Connectivity-based parcellation of broca's area. *Cereb. Cortex* 17, 816–825. doi: 10.1093/cercor/bhk034

Baldo, J. V., Arevalo, A., Patterson, J. P., and Dronkers, N. F. (2013). Grey and white matter correlates of picture naming: evidence from a voxel-based lesion analysis of the boston naming test. *Cortex* 49, 658–667. doi: 10.1016/j.cortex.2012.03.001

Bavelier, D., Corina, D., Jezzard, P., Clark, V., Karni, A., Lalwani, A., et al. (1998). Hemispheric specialization for English and ASL: left invariance-right variability. *Neuroreport* 9, 1537–1542. doi: 10.1097/00001756-199805110-00054

Beauchamp, M. S. (2005). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr. Opin. Neurobiol.* 15, 145–153. doi: 10.1016/j.conb.2005.03.011

Beauchamp, M. S., Yasar, N. E., Frye, R. E., and Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *Neuroimage* 41, 1011–1020. doi: 10.1016/j.neuroimage.2008.03.015

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–312. doi: 10.1038/35002078

Binder, J. R., and Desai, R. H. (2011). The neurobiology of semantic memory. *Trends Cogn. Sci.* 15, 527–536. doi: 10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055

Boatman, D., Gordon, B., Hart, J., Selnes, O., Miglioretti, D., and Lenz, F. (2000). Transcortical sensory aphasia: revisited and revised. *Brain* 123 (Pt. 8), 1634–1642. doi: 10.1093/brain/123.8.1634

Bonner, M. F., Peelle, J. E., Cook, P. A., and Grossman, M. (2013). Heteromodal conceptual processing in the angular gyrus. *Neuroimage* 71, 175–186. doi: 10.1016/j.neuroimage.2013.01.006

Buchsbaum, B. R., and D'Esposito, M. (2009). Repetition suppression and reactivation in auditory-verbal short-term recognition memory. *Cereb. Cortex* 19, 1474–1485. doi: 10.1093/cercor/bhn186

Buchsbaum, B. R., Olsen, R. K., Koch, P., and Berman, K. F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron* 48, 687–697. doi: 10.1016/j.neuron.2005.09.029

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., et al. (2009). Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *J. Neurosci.* 29, 1860–1873. doi: 10.1523/JNEUROSCI.5062-08.2009

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

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., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage* 14, 427–438. doi: 10.1006/nimg.2001.0812

Caplan, D., Waters, G. S., and Hildebrandt, N. (1997). Determinants of sentence comprehension in aphasic patients in sentence-picture matching tasks. *J. Speech Lang. Hear. Res.* 40, 542–555. doi: 10.1044/jslhr.4003.542

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., and Zilles, K. (2006). The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. *Neuroimage* 33, 430–448. doi: 10.1016/j.neuroimage.2006.06.054

Catani, M., Jones, D. K., and ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Ann. Neurol.* 57, 8–16. doi: 10.1002/ana.20319

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9, 179–194. doi: 10.1006/nimg.1998.0395

Damasio, A. R. (1989). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33, 25–62. doi: 10.1016/0010-0277(89)90005-X

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

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

DeWitt, 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

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D. Jr., Redfern, B. B., and Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92, 145–177. doi: 10.1016/j.cognition.2003.11.002

Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., and Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *Neuroimage* 59, 2349–2361. doi: 10.1016/j.neuroimage.2011.09.017

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., and Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* 30, 2907–2926. doi: 10.1002/hbm.20718

Frey, S., Campbell, J. S., Pike, G. B., and Petrides, M. (2008). Dissociating the human language pathways with high angular resolution diffusion fiber tractography. *J. Neurosci.* 28, 11435–11444. doi: 10.1523/JNEUROSCI.2388-08.2008

Gilaie-Dotan, S., Kanai, R., Bahrami, B., Rees, G., and Saygin, A. P. (2013). Neuroanatomical correlates of biological motion detection. *Neuropsychologia* 51, 457–463. doi: 10.1016/j.neuropsychologia.2012.11.027

Golestani, N., and Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Grosbras, M. H., Beaton, S., and Eickhoff, S. B. (2012). Brain regions involved in human movement perception: a quantitative voxel-based meta-analysis. *Hum. Brain Mapp.* 33, 431–454. doi: 10.1002/hbm.21222

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301. doi: 10.1016/j.bandl.2005.06.001

Hein, G., and Knight, R. T. (2008). Superior temporal sulcus—It's my area: or is it? *J. Cogn. Neurosci.* 20, 2125–2136. doi: 10.1162/jocn.2008.20148

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

Humphries, C. J., Sabri, M., Heugel, N., Lewis, K., and Liebenthal, E. (2013). Pattern specific adaptation to speech and non-speech sounds in human auditory cortex. *Soc. Neurosci.* Abst. 354.21/SS7

Kanai, R., Bahrami, B., Duchaine, B., Janik, A., Banissy, M. J., and Rees, G. (2012). Brain structure links loneliness to social perception. *Curr. Biol.* 22, 1975–1979. doi: 10.1016/j.cub.2012.08.045

Kilian-Hutten, N., Valente, G., Vroomen, J., and Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *J. Neurosci.* 31, 1715–1720. doi: 10.1523/JNEUROSCI.4572-10.2011

Kosslyn, S. M., Shin, L. M., Thompson, W. L., McNally, R. J., Rauch, S. L., Pitman, R. K., et al. (1996). Neural effects of visualizing and perceiving aversive stimuli: a PET investigation. *Neuroreport* 7, 1569–1576. doi: 10.1097/00001756-199607080-00007

Laird, A. R., Lancaster, J. L., and Fox, P. T. (2005). BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* 3, 65–78. doi: 10.1385/NI:3:1:065

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Lee, K. H., and Siegle, G. J. (2012). Common and distinct brain networks underlying explicit emotional evaluation: a meta-analytic study. *Soc. Cogn. Affect. Neurosci.* 7, 521–534. doi: 10.1093/scan/nsp001

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., and Desai, A. (2013). Neural dynamics of phonological processing in the dorsal auditory stream. *J. Neurosci.* 33, 15414–15424. doi: 10.1523/JNEUROSCI.1511-13.2013

MacSweeney, M., Campbell, R., Woll, B., Brammer, M. J., Giampietro, V., David, A. S., et al. (2006). Lexical and sentential processing in british sign language. *Hum. Brain Mapp.* 27, 63–76. doi: 10.1002/hbm.20167

Mesulam, M. M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Ann. Neurol.* 28, 597–613. doi: 10.1002/ana.410280502

Mesulam, M. M. (1998). From sensation to cognition. *Brain* 121(Pt. 6), 1013–1052. doi: 10.1093/brain/121.6.1013

Meyer, K., and Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends Neurosci.* 32, 376–382. doi: 10.1016/j.tins.2009.04.002

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

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cereb. Cortex* 17, 2251–2257. doi: 10.1093/cercor/bhl133

Ochiai, T., Grimault, S., Scavarda, D., Roch, G., Hori, T., Riviere, D., et al. (2004). Sulcal pattern and morphology of the superior temporal sulcus. *Neuroimage* 22, 706–719. doi: 10.1016/j.neuroimage.2004.01.023

Parker, G. J., Luzzi, S., Alexander, D. C., Wheeler-Kingshott, C. A., Ciccarelli, O., and Lambon Ralph, M. A. (2005). Lateralization of ventral and dorsal auditory-language pathways in the human brain. *Neuroimage* 24, 656–666. doi: 10.1016/j.neuroimage.2004.08.047

Paulesu, E., Frith, C. D., and Frackowiak, R. S. (1993). The neural correlates of the verbal component of working memory. *Nature* 362, 342–345. doi: 10.1038/362342a0

Planton, S., Jucla, M., Roux, F. E., and Demonet, J. F. (2013). The "handwriting brain": a meta-analysis of neuroimaging studies of motor versus orthographic processes. *Cortex* 49, 2772–2787. doi: 10.1016/j.cortex.2013.05.011

Price, C. J. (2000). The anatomy of language: contributions from functional neuroimaging. *J. Anat.* 197(Pt. 3), 335–359. doi: 10.1046/j.1469-7580.2000.19730335.x

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* 1191, 62–88. doi: 10.1111/j.1749-6632.2010.05444.x

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

Purcell, J. J., Turkeltaub, P. E., Eden, G. F., and Rapp, B. (2011). Examining the central and peripheral processes of written word production through meta-analysis. *Front. Psychol.* 2:239. doi: 10.3389/fpsyg.2011.00239

Raizada, R. D., and Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron* 56, 726–740. doi: 10.1016/j.neuron.2007.11.001

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

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

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., et al. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18035–18040. doi: 10.1073/pnas.0805234105

Segal, E., and Petrides, M. (2012). The morphology and variability of the caudal rami of the superior temporal sulcus. *Eur. J. Neurosci.* 36, 2035–2053. doi: 10.1111/j.1460-9568.2012.08109.x

Sowell, E. R., Thompson, P. M., Rex, D., Kornsand, D., Tessner, K. D., Jernigan, T. L., et al. (2002). Mapping sulcal pattern asymmetry and local cortical surface gray matter distribution *in vivo*: maturation in perisylvian cortices. *Cereb. Cortex* 12, 17–26. doi: 10.1093/cercor/12.1.17

Talairach, J., and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain*. New York, NY: Thieme Medical Publishers.

Turkeltaub, P. E., and Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain Lang.* 114, 1–15. doi: 10.1016/j.bandl.2010.03.008

Turkeltaub, P. E., Eden, G. F., Jones, K. M., and Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16, 765–780. doi: 10.1006/nimg.2002.1131

Turken, A. U., and Dronkers, N. F. (2011). The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Front. Syst. Neurosci.* 5:1. doi: 10.3389/fnsys.2011.00001

Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., and Anderson, C. H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. *J. Am. Med. Inform. Assoc.* 8, 443–459. doi: 10.1136/jamia.2001.0080443

Van Overwalle, F., and Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage* 48, 564–584. doi: 10.1016/j.neuroimage.2009.06.009

Ward, B. D. (2000). *Simultaneous Infererence for fMRI Data*. Available online at: http://afni.nimh.nih.gov/pub/dist/doc/manual/AlphaSim.pdf

Wise, R. J., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., and Warburton, E. A. (2001). Separate neural subsystems within 'Wernicke's area'. *Brain* 124, 83–95. doi: 10.1093/brain/124.1.83

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 June 2014; accepted: 26 August 2014; published online: 11 September 2014.

Citation: Liebenthal E, Desai RH, Humphries C, Sabri M and Desai A (2014) The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. Front. Neurosci. ***8****:289. doi: 10.3389/fnins.2014.00289*

This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

REVIEW ARTICLE
published: 01 December 2014
doi: 10.3389/fnins.2014.00386

Neural pathways for visual speech perception

Lynne E. Bernstein[1]* and Einat Liebenthal[2,3]

[1] *Department of Speech and Hearing Sciences, George Washington University, Washington, DC, USA*
[2] *Department of Neurology, Medical College of Wisconsin, Milwaukee, WI, USA*
[3] *Department of Psychiatry, Brigham and Women's Hospital, Boston, MA, USA*

Edited by:
Josef P. Rauschecker, Georgetown University School of Medicine, USA

Reviewed by:
Ruth Campbell, University College London, UK
Josef P. Rauschecker, Georgetown University School of Medicine, USA
Kaisa Tiippana, University of Helsinki, Finland

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

This paper examines the questions, what levels of speech can be perceived visually, and how is visual speech represented by the brain? Review of the literature leads to the conclusions that every level of psycholinguistic speech structure (i.e., phonetic features, phonemes, syllables, words, and prosody) can be perceived visually, although individuals differ in their abilities to do so; and that there are visual modality-specific representations of speech *qua* speech in higher-level vision brain areas. That is, the visual system represents the modal patterns of visual speech. The suggestion that the auditory speech pathway receives and represents visual speech is examined in light of neuroimaging evidence on the auditory speech pathways. We outline the generally agreed-upon organization of the visual ventral and dorsal pathways and examine several types of visual processing that might be related to speech through those pathways, specifically, face and body, orthography, and sign language processing. In this context, we examine the visual speech processing literature, which reveals widespread diverse patterns of activity in posterior temporal cortices in response to visual speech stimuli. We outline a model of the visual and auditory speech pathways and make several suggestions: (1) The visual perception of speech relies on visual pathway representations of speech *qua* speech. (2) A proposed site of these representations, the temporal visual speech area (TVSA) has been demonstrated in posterior temporal cortex, ventral and posterior to multisensory posterior superior temporal sulcus (pSTS). (3) Given that visual speech has dynamic and configural features, its representations in feedforward visual pathways are expected to integrate these features, possibly in TVSA.

Keywords: functional organization, audiovisual processing, speech perception, lipreading, visual processing

INTRODUCTION

This paper examines the questions, what levels of speech can be perceived visually, and how is visual speech represented by the brain? These questions would hardly have arisen 50 years ago. Mid-twentieth century speech perception theories were strongly influenced by the expectation that speech perception is an *auditory* function for processing *acoustic* speech stimuli (Klatt, 1979; Stevens, 1981), perhaps, in close coordination with the motor system (Liberman et al., 1967; Liberman, 1982). At the time, theorizing about speech perception was unrelated to evidence about visual speech perception (lipreading[1]), even though there were reports available in the literature showing that speech can be perceived visually. For example, there was extensive evidence during most of the twentieth century that lipreading can substitute for hearing in the education of deaf children (Jeffers and Barley, 1971), and there was evidence about the important role of lipreading in combination with residual hearing for children and adults with hearing impairments (Erber, 1971). The basic finding in normal-hearing adults that vision can compensate for hearing under noisy conditions was reported by mid-twentieth century (Sumby and Pollack, 1954). Even the report by McGurk and MacDonald (1976) that a visual speech stimulus mismatched with an auditory stimulus can alter perception of an auditory speech stimulus, an effect that has come to be known as the McGurk effect, had few responses in the literature until a number of years following its publication.

Research efforts to explain the McGurk effect and understand its general implications for speech perception and multisensory processing began in the 1980s (e.g., Massaro and Cohen, 1983; Liberman and Mattingly, 1985; Campbell et al., 1986; Green and Kuhl, 1989), as did forays into theoretical explanations for how auditory and visual speech information combines perceptually (Liberman and Mattingly, 1985; Massaro, 1987; Summerfield, 1987). In the following decade, in tandem with the development of new neuroimaging technologies, reports emerged that visual speech stimuli elicit auditory cortical responses (Sams et al., 1991; Calvert et al., 1997), results that seemed consistent with the phenomenal experience of the McGurk effect as a change in the auditory perception of speech. In the 1990s, breakthrough

[1]The term *lipreading* is used in this paper to refer to perceiving speech by vision. An alternate term that appears in the literature is *speechreading*. This term is sometimes used to emphasize the point that visual speech perception is more than perception of lips, and sometimes it is used to refer to visual speech perception augmented by residual hearing in individuals with hearing impairments.

research on multisensory processing in cat superior colliculus was presented by Stein and Meredith (1993). Their evidence about multisensory neuronal integration provided a potential neural mechanism for explaining how auditory and visual speech information is processed (Calvert, 2001), specifically, that auditory and visual speech information converges early in the stream of processing.

Evidence for multisensory inputs to classically defined unisensory cortical areas (e.g., Falchier et al., 2002; Foxe et al., 2002) helped to shift the view of the sensory pathways as modality-specific until the levels of association cortex (Mesulam, 1998) toward the view that the brain is massively multisensory (Foxe and Schroeder, 2005; Ghazanfar and Schroeder, 2006). Findings suggesting the possibility that visual speech stimuli have special access to the early auditory speech processing pathway (Calvert et al., 1997; Ludman et al., 2000; Pekkola et al., 2005) were consistent with the emerging multisensory view. More recently, reconsideration of the motor theory of speech perception (Liberman and Mattingly, 1985) and mirror neuron system theory (Rizzolatti and Arbib, 1998; Rizzolatti and Craighero, 2004) have led inquiry into the role of somatomotor processing in speech perception, including visual speech perception (Hasson et al., 2007; Skipper et al., 2007a; Matchin et al., 2014). In this context, a question has been the extent to which visual speech is represented in frontal cortex (Callan et al., 2014). Thus, both the auditory and somatomotor systems have been studied for their roles in representing visual speech.

Curiously, the role of the visual system in representing speech has received less attention than the role of the auditory speech pathways. What is particularly curious is that the visual speech stimulus is psycholinguistically extremely rich, as shown below, yet there has been little research that has focused on how the visual system represents visible psycholinguistic structure (i.e., phonetic features, phonemes, syllables, prosody, and even words); although there have been, as we discuss below, multiple studies that show that speech activates areas in high-level visual pathways (for reviews, Campbell, 2008, 2011). The absence of pointed investigations of how visual speech is represented—in contrast to the detailed knowledge about auditory speech representations—is surprising, because sensory systems transduce specific types of energy such as light and sound, each affording its own form of evidence about the environment, including speech; and the current view of multisensory interactions does not overturn the classical hierarchical models of auditory and visual sensory pathways (e.g., Felleman and Van Essen, 1991; Kaas and Hackett, 2000; Rauschecker and Tian, 2000) as much as it enriches them. Clearly, the diverse evidence for multisensory interactions needs to be reconciled with evidence pointing to modality-specific stimulus representations and processing (Hertz and Amedi, 2014). This review explores the expectation that perception of visual speech stimuli requires visual representations of the stimuli through the visual pathways.

In this paper, we review the visual speech perception literature to support the view that every psycholinguistic level of speech organization is visible. That being the case, we consider the cortical representation of auditory speech as a possible model for the organization of visual speech processing. We suggest that research on the auditory organization of speech processing does not in fact encourage the notion that visual speech perception can be explained by multisensory connections alone. We propose a model that posits modality-specific as well as amodal speech processing pathways. **Figure 1** summarizes our model, which is discussed in detail further below.

VISUAL SPEECH PERCEPTION

IMPLICATIONS OF INDIVIDUAL DIFFERENCES IN LIPREADING ABILITY

Any discussion of visual speech perception and its underlying neural mechanisms needs to acknowledge the fact of large inter-individual variation, both within and across normal-hearing and deaf populations (Bernstein et al., 2000, 2001; Auer and Bernstein, 2007; Tye-Murray et al., 2014). The differences are so large that findings on visual speech processing can probably not be accurately interpreted without knowing something about individual participants' lipreading ability and auditory experience.

For example, in a test of words correctly lipread in isolated sentences, the scores by deaf lipreaders ranged from zero to greater than 85% correct (Bernstein et al., 2000). Deaf lipreaders were able to identify as many as 42% of isolated monosyllabic words from a list of highly confusable rhyming words (each test word rhymed with five other English words). Among adults with normal hearing, there was a narrower performance range for the same stimulus materials: There were individuals with scores as low as zero and ones with very good lipreading ability with scores as high as 75% correct words in sentences and 24% correct on the isolated rhyming words. Analyses of phoneme confusions in lipreading sentences suggested that the deaf participants were using more visual phonetic feature information than the hearing adults. But the individual variation in lipreading sentences accounted for by isolated word vs. isolated phoneme identification (using non-sense syllables) scores showed that isolated words accounted for more variance than phonemes: Word identification scores with isolated rhyme words accounted for between 66 and 71% of the variance in words-in-sentences scores for deaf lipreaders and between 44 and 64% of the variance for normal-hearing lipreaders, values commensurate with other reports (Conklin, 1917; Utley, 1946; Lyxell et al., 1993). In Bernstein et al. (2000), phoneme identification in non-sense syllables accounted for between 21 and 43% of the variance in words-in-sentences scores for deaf lipreaders and between 6 and 18% of the variance for normal-hearing lipreaders. When regression was used to predict words-in-sentences scores, only participant group (deaf, normal-hearing) and isolated word scores were significant predictors (multiple R between 0.88 and 0.90). Additional studies confirm that the best lipreaders experienced profound congenital hearing loss, but that even among normal-hearing adults there are individuals with considerable lipreading expertise (Mohammed et al., 2006; Auer and Bernstein, 2007).

Individuals with hearing impairments may rely primarily on visual speech, even in the context of hearing aid and cochlear implant usage (Rouger et al., 2007; Bernstein et al., 2014; Bottari et al., 2014; Song et al., 2014). Lipreading ability in individuals with hearing loss, including those with congenital impairments is likely associated with a wide range of neuroplastic effects, including take-over of auditory processing areas by vision (Karns et al.,

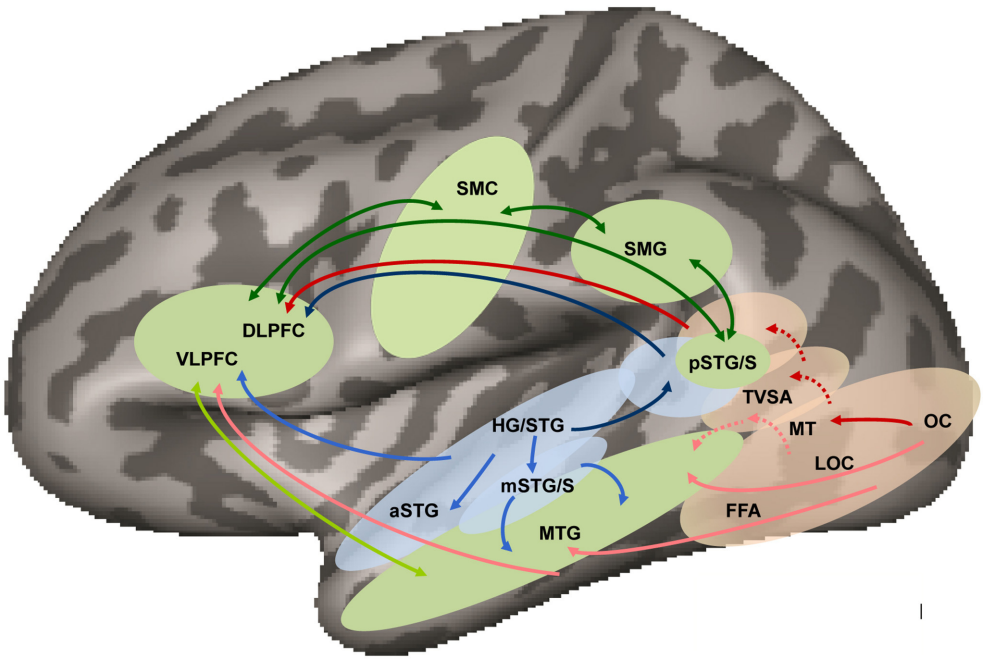


FIGURE 1 | Neuroanatomical working model of audiovisual speech perception in the left hemisphere based on models of dual visual (Wilson et al., 1993; Haxby et al., 1994; Ungerleider et al., 1998; Weiner and Grill-Spector, 2013) and auditory (Romanski et al., 1999; Hickok and Poeppel, 2007; Saur et al., 2008; Rauschecker and Scott, 2009; Liebenthal et al., 2010) pathways and audiovisual integration (Beauchamp et al., 2004) in humans. Audiovisual speech is processed in auditory (blue) and visual (pink) areas projecting to amodal (green) middle temporal cortex via auditory (light blue arrows) and visual (light red arrows) ventral pathways terminating in VLPFC, and to multimodal posterior temporal cortex via auditory (dark blue) and visual (dark red) dorsal pathways terminating in DLPFC. Specialization for phoneme processing is suggested to exist in both auditory and visual pathways, at the level of mSTG/S and TVSA, respectively, although the pattern of connectivity of TVSA (shown in red dotted arrows), and whether it is part of the ventral and/or dorsal visual streams is unknown. Multimodal or amodal areas in the ventral and dorsal streams connect bi-directionally via direct and indirect ventral (light green arrows) and dorsal (dark green arrows) pathways. (HG/STG, Heschl's gyrus/superior temporal gyrus; aSTG, anterior superior temporal gyrus; mSTG/S, middle superior temporal gyrus and sulcus; pSTG/S, posterior superior temporal gyrus and sulcus; MTG, middle temporal gyrus; OC, occipital cortex; FFA, fusiform face area; LOC, lateral occipital complex; MT, middle temporal area; TVSA, temporal visual speech area; SMG, supramarginal gyrus; SMC, somatomotor cortex; VLPFC, ventrolateral prefrontal cortex; DLPFC, dorsolateral prefrontal cortex).

2012; Bottari et al., 2014) or somatosensation (Levanen et al., 1998; Auer et al., 2007; Karns et al., 2012), and alterations of sub-cortical connections (Lyness et al., 2014).

VISIBLE LEVELS OF SPEECH

From a psycholinguistic perspective, speech has a hierarchical structure comprising features, phonemes, syllables, words, phrases, and larger units such as utterances, sentences, and discourse. The questions here are which of these levels can be perceived visually, and whether any type of these speech patterns is represented in visual modality-specific areas. As with auditory speech perception, we expect that at a minimum visual speech perception extends to the physical properties of speech, that is, its *phonetic* feature properties, and that those properties express the vowels, consonants, and prosody of a language. The term *phonemic* refers to language-specific segmental (vowel and consonant) properties. Thus, for example, the term *phonetic* applies to speech features without necessarily specifying a particular language, and *phonemic* refers to segmental distinctions used by a particular language to distinguish among words (Catford, 1977). Prosody comprises phonetic attributes that span words or phrases, such as lexical stress in English (e.g., the distinction between the verb in "to re*cord*" and the noun in "the *re*cord"), and intonation (e.g., pronunciation of the same phrase as an exclamation or a statement, "we won!/?"). Necessarily, physical acoustic phonetic speech signals are different than optical phonetic speech signals; and although they may convey the same linguistic content, they are expected to be represented initially by different peripheral, subcortical, and primary sensory areas that code different low-level basic sensory features (e.g., light intensities vs. sound intensities, spatio-temporal vs. temporal frequencies, etc.). As we suggest below, there is the possibility that modality-specific representations exist to the level of whole words. But we do not expect separate representations of the meanings of individual words or of whole visual multi-word utterances, although there may be highly frequent utterances that are represented as such.

FEATURES, PHONEMES, AND VISEMES

Speech production simultaneously produces the sounds and sights of speech, but the vocal tract shapes, glottal vibrations, and velar gestures that produce acoustic speech (Stevens, 1998) are not all directly visible. Some of them are visible as correlated motions

of the jaw and the cheeks (Yehia et al., 1998; Jiang et al., 2002, 2007). An ongoing idea in the literature is that visual speech is too impoverished to convey much phonetic information (Kuhl and Meltzoff, 1988). This idea is supported by examples of poor lipreading performance and by focusing on how acoustic signals are generated. For example, the voicing feature (i.e., the feature that distinguishes "b" from "p") is typically expressed acoustically in pre-vocalic position in terms of glottal vibration characteristics such as onset time (Lisker et al., 1977). But the glottis is not a visible structure, so a possible inference is that the voicing feature cannot be perceived visually. However, there are other phonetic attributes that contribute to voicing distinctions. For example, post-vocalic consonant voicing depends greatly on vowel duration (Raphael, 1971), and vowel duration—the duration of the open mouth gesture—is visible. When visual consonant identification was compared across initial (C[=consonant]V[=vowel]), medial (VCV), and final (VC) position (Van Son et al., 1994), identification of final consonants was 44% correct in contrast to 28% for consonants elsewhere. The point is that both optical and acoustic phonetic attributes instantiate speech features on the basis of diverse sensory information; so the visibility of speech features or phonemes cannot be inferred accurately from a simple one-to-one mapping between the visibility of speech production anatomy (e.g., lips, mouth, tongue, glottis) and speech features (e.g., voicing, place, manner, nasality).

At the same time, the reduction in visual vs. auditory speech information needs to be taken into account. The concept of the *viseme* was invented to describe and account for the somewhat stable patterns of lipreaders' phoneme confusions (Woodward and Barber, 1960; Fisher, 1968; Owens and Blazek, 1985). Visemes are sets such as /p, b, m/ that are typically formed using some grouping principle such as hierarchical clustering of consonant confusions from phoneme identification paradigms (Walden et al., 1977; Auer and Bernstein, 1997; Iverson et al., 1998). A typical rule is on the order of grouping together phonemes whose mutual confusions account for around 70% of responses. Massaro suggested that, "Because of the data-limited property of visible speech in comparison to audible speech, many phonemes are virtually indistinguishable by sight, even from a natural face, and so are expected to be easily confused" (p. 316); and that, "a difference between visemes is significant, informative, and categorical to the perceiver; a difference within a viseme class is not" (Massaro et al., 2012, p. 316).

However, most research that has used the viseme concept has involved phoneme identification tasks, for which there is a need to account for identification errors. A difference within a viseme class could be significant and informative. It could also be categorical at the level of a feature. Indeed, when presented with pairs of spoken words that differed only in terms of phonemes from within putative viseme sets, participants (deaf and normal-hearing adults) were able to identify which of the spoken words corresponded to an orthographic target word (Bernstein, 2012). That is, each word pair in the target identification paradigm was constructed so that in sequential order each of its phonemes was selected from within the same viseme. The visemes were defined along the standard lines of constructing viseme sets. An additional set of word pairs was constructed from within sets that comprised even higher levels of confusability than used to construct visemes (referred to as "phoneme equivalence classes"; Auer and Bernstein, 1997). Normal-hearing lipreaders with above-average lipreading scored between 65 and 80% correct word identification with stimuli comprising the *subvisemic* phoneme sets (i.e., the sets of very similar phonemes). Deaf participants scored between 80 and 100% correct on those word-pairs. This would not have been possible if the phonemes that comprise visemes were not significant or informative. Thus, while there is no doubt that visual speech stimuli afford reduced phonetic detail in support of phoneme categories, there is also evidence that perceivers are not limited to perceiving viseme categories.

Interestingly, not only are perceivers able to perceive speech stimuli based on fine visual phonetic distinctions, they are also able to make judgments of the reliability of their own perceptions, apparently in terms of perceived phoneme or feature stimulus-to-response discrepancies. In a study of sentence lipreading (Demorest and Bernstein, 1997), deaf and normal-hearing adults were presented with isolated spoken sentences for open set identification of the words in the sentences. Participants were asked to type what they thought the talker had said and also to rate their confidence in their typed responses, and they received no feedback on their performance. Confidence ratings ranged from 0 = "no confidence—I guessed" to 7 = "complete confidence—I understood every word." Scoring for how well sentences were lipread included a measure of the perceptual distance based on phoneme alignments between the stimulus and the response and was computed using a sequence comparison algorithm (Kruskal and Wish, 1978; Bernstein et al., 1994) that aligned stimulus and response phoneme sequences using visual perceptual phoneme dissimilarity weights. As an example, when the stimulus sentence was, "Why should I get up so early in the morning?" and the response was, "Watch what I'm doing in the morning," casual inspection of the stimulus and response suggest that they have similar phoneme strings even when some of the words were incorrectly identified. The sequence comparator aligned the phonemes of these two sentences as follows (in Arpabet phonemic notation):

```
Stimulus: wA SUd A gEt ^p so Rli In Dx morn|G
Response: wa C-- - wxt Am du |G- In Dx morn|G
```

Perusal of the string alignment suggests that there were phoneme similarities even when whole words were incorrect. A visual distance score was computed for each stimulus-response pair based only on the distances between aligned *incorrect* phonemes (e.g., "S" vs. "C" in the example) normalized by stimulus length in phonemes. Correct phonemes did not contribute to distance scores. Correlations between stimulus-response distances and subjective confidence ratings showed that as stimulus-response distance (perceptual dissimilarity) increased, subjective confidence went down (reliable Pearson correlations of −0.511 for normal-hearing and −0.626 for deaf). These findings suggest that deaf and hearing adults have access to perceptual representations that preserve to some extent the phonetic information in the visual stimulus and thereby allow them to judge discrepancy between the stimulus and their own response. Thus, both

this approach and the target identification approach described above reveal that sub-visemic speech information is significant and informative.

If lipreading relies on visual image processing, there should be direct relationships between the structure of the visual images and perception. A study (Jiang et al., 2007) addressed the relationship between optical recordings and visual speech perception. Recordings were made of 3-dimensional movement of the face and simultaneous video while talkers produced many different CV syllables (i.e., all the initial English consonants, followed by one of three different vowels, and spoken by four different talkers). If visual stimuli drive visual speech perception, than there should be a second-order isomorphism (Shepard and Chipman, 1970) between optical data and perception such that the dissimilarity of physical speech signals should map onto perceptual dissimilarity. The study showed that a linearly warped physical stimulus dissimilarity space was highly effective in accounting for the perceptual structure of phoneme identification for spoken CVs. Across talkers, the 3-dimensional face movement data accounted for between 46 and 66% of the variance in perceptual dissimilarities among CV stimuli.

SPOKEN WORDS

Visual spoken word recognition has been studied in experiments that were designed to investigate the pattern of visual confusions among spoken words. These studies show that visual dissimilarities affect perception to the level of spoken word identification.

For example, Mattys et al. (2002) presented isolated mono- and disyllabic spoken word stimuli to normal-hearing and deaf lipreaders for open-set visual identification. The words were selected so that they varied in terms of the number of words in the lexicon with which each was potentially confusable based on visual phoneme confusability (Iverson et al., 1998). The results showed that visual phoneme confusability predicted the relative accuracy levels for word identification by both participant groups, and phoneme errors tended to be from within groups of visually more confusable phonemes.

Auer (2002) visually presented isolated spoken monosyllabic words to deaf and normal-hearing lipreaders and modeled perception using auditory vs. visual phoneme confusion data. The visual confusions were better predictors of visual spoken word recognition than auditory confusions. Strand and Sommers (2011) followed up and tested monosyllabic words in visual-only and auditory-only (with noise background) conditions. They modeled lexical competition effects separately for visual vs. auditory phoneme similarity and showed that measures of similarity (i.e., lexical competition) that were based on one modality were not good predictors of word identification accuracy for the other modality.

PROSODY

Prosody comprises stress and intonation (Risberg and Lubker, 1978; Jesse and McQueen, 2014). Several studies have investigated visual prosody perception in normal-hearing adults (Fisher, 1969; Lansing and McConkie, 1999; Scarborough et al., 2007; Jesse and McQueen, 2014). Results suggest that prosody is perceived visually.

For example, emphatic stress for specific words such as, "We OWE you a yoyo," vs., "We owe YOU a yoyo," was perceived quite accurately (70%, chance = 33.3%), while perception of whether those sentences were spoken as statements or questions was perceived somewhat less accurately (60%, chance = 50%) (Bernstein et al., 1989; see also, Lansing and McConkie, 1999). Lexical stress in bisyllabic words such as *SUBject* (the noun) and *subJECT* (the verb) can be visually discriminated (62%, chance = 50%), as can phrasal stress that distinguishes (in sentences with stress on one of the names in "So, [name1] gave/sang [name2] a song from/by [name3]") (54% correct, chance = 25%) (Scarborough et al., 2007). In the latter study, larger and faster face movements were associated with the perception of stress. For example, lower lip opening peak velocity and the size of lip opening were related to lexical stress perception.

Even whole head movement has been shown to be correlated with prosody (63% of variance accounted for between voice pitch and six components of head movement) (Munhall et al., 2004), with head movement contributing to the accuracy of speech perception in noise. Visible head movement can be used by talkers for perceiving emphasis (Lansing and McConkie, 1999).

Visual prosody perception has been studied in infants. Prosody is used in parsing connected speech and may thereby assist infants in acquiring their native language (Johnson et al., 2014). Visible prosody is likely a contributor to infants' demonstrated sensitivity to language differences in visual speech stimuli (Weikum et al., 2007).

INTERIM SUMMARY

In answer to our question, What levels of speech can be perceived visually? we conclude that all levels of speech patterns (from features to connected speech) that can be heard can also be visually perceived, at least by the more skilled of lipreaders. Visual phoneme categories have internal perceptual structure that is different from that of auditory phoneme categories. At least in the better lipreaders, there may be visual modality-specific syllable or word pattern representations. Research on visual prosody suggests that it can be perceived in multisyllabic words and in connected speech. Thus, the perceptual evidence is fully compatible with the possibility that the visual speech perception relies on extensive visual modality-specific neural representations.

AN AUDITORY REPRESENTATION OF VISUAL SPEECH?

The earliest human neuroimaging studies on lipreading revealed activity in the region of primary auditory cortex, leading to discussions about the role of the auditory pathway in processing visual speech, perhaps as early as the primary auditory cortex (Sams et al., 1991; Calvert et al., 1997). Interpretations of the observed activity pointed to a role for the auditory pathway akin to its role in processing auditory speech stimuli: For example, "results show that visual information from articulatory movements has an entry into the auditory cortex" (Sams et al., 1991); "activation of primary auditory cortex during lipreading suggests that these visual cues may influence the perception of heard speech before speech sounds are categorized in auditory association cortex into distinct phonemes" (Calvert et al., 1997); "Visual speech has access to auditory sensory memory" (Möttönen et al.,

2002); and "seen speech with normal time-varying characteristics appears to have preferential access to 'purely' auditory processing regions specialized for language" (Calvert and Campbell, 2003).

These statements were not accompanied by an explicit model or theory about how visual speech stimuli are represented by visual cortical areas upstream of auditory cortex. One reading of these statements is that rather than computing the patterns of visual speech *qua* speech within the visual system, there is a special route for visual speech to the auditory pathway where it is represented as though it were an auditory speech stimulus.

Alternatively, visual speech patterns are integrated somehow within the visual system and then projected to the primary auditory cortex where they are re-represented. However, the re-representation of information is considered to be a computationally untenable solution for the brain (von der Malsburg, 1995).

Another possibility is that visual stimuli are analyzed by the visual system only to the level of features such as motion or edges that are not integrated specifically as speech, and those feature representations are projected to the auditory pathway. But then it would be necessary to explain at what point the unbound information specific to speech was recognized as speech and was prioritized for entry into the auditory pathway. This possibility clearly suggests a "chicken and egg" problem.

Whatever its implications, there have been various attempts to confirm with neuroimaging in the human that primary auditory cortex activation levels increase following visual speech stimuli, with mixed results (Ludman et al., 2000; Bernstein et al., 2002; Calvert and Campbell, 2003; Besle et al., 2004; Pekkola et al., 2005; Okada et al., 2013). However, were visual speech prioritized for entry to auditory cortex, we might expect to see its effects more consistently.

Even when obtained, higher activation levels measured in the region of primary auditory cortex are of course not unambiguous with regard to the underlying neural response. They could for example be due to auditory imagery (Hickok et al., 2003). Or visual motion could drive the response (Okada et al., 2013). The location of primary auditory cortex could be inaccurately identified, particularly with group averaging, as non-invasive methods are imprecise in delineating the auditory core vs. belt cortex (Desai et al., 2005). Finally, a definite possibility is that activity measured with functional imaging in the region of the auditory cortex is attributable to feedback rather than visual stimulus pattern representation (Calvert et al., 2000; Schroeder et al., 2008).

There are relevant monkey data concerning the representation of input across modalities. Direct connections have been demonstrated from auditory core and parabelt to V1 in monkeys (Falchier et al., 2002) and from V2 to caudal auditory cortex (Falchier et al., 2010). These studies did not show connections from V1 to A1. The character of the connections is that of feedback through the dorsal visual pathway, commensurate with the function of representing extra-personal peripheral space and motion. "These results suggest a model in which putative unisensory visual and auditory cortices do not interact in a classical feedforward–feedback relationship but rather by way of a feedback loop. A possible implication of this organization is that the dominant effects of these connections between early sensory areas are modulatory" (Falchier et al., 2010). Importantly, monkey work has also shown that visual stimuli can modulate auditory responses in primary and secondary auditory fields *independent of the visual stimulus categories* (Kayser et al., 2008), and similar findings have been generalized to modulation of auditory cortices by somatosensory stimuli (Lemus et al., 2010). Thus, while there are functional connections, these connections between early sensory areas may serve primarily downstream modulatory functions and not upstream representation of perceptual detail needed for recognizing stimulus categories.

Overall, replication of primary auditory cortex activation by visual speech has not been completely successful, explanations invoking phonetic processing have been vague with regard to upstream visual input computations, and animal research has not been supportive of the possibility that visual speech perception is the result of representing the visual speech information through activation of auditory speech representations. The research on auditory speech processing, to which we now turn, also discourages notions about the representation of visual speech by the auditory pathway.

THE AUDITORY REPRESENTATION OF SPEECH

The research on auditory speech processing is fairly clear in establishing that phonetic and phonemic speech representations in superior temporal regions beyond auditory core are viewed as modal, that is, abstracted from low-level acoustic characteristics but preserving some of their attributes. These modality specific auditory representations are not predicted to also respond to visual speech stimulus phonetic features or phonemes. Thus, our neuroanatomical model in **Figure 1** posits distinct visual and auditory pathways to the level of pSTS.

Emerging work in the human suggests that neurons in the left superior temporal gyrus (STG) show selectivity to spectrotemporal acoustic cues that map to distinct phonetic features (e.g., manner of articulation) and not to distinct phonemes. Sensitivity to different phonetic features has been demonstrated in the middle and posterior STG using data-mining algorithms to identify patterns of activity in functional magnetic resonance imaging (fMRI) (Formisano et al., 2008; Kilian-Hutten et al., 2011; Humphries et al., 2013) and in intracranial (Chang et al., 2010; Steinschneider et al., 2011; Chan et al., 2014; Mesgarani et al., 2014) responses. There is now also conclusive evidence that an area in the left middle and ventral portion of STG and adjacent superior temporal sulcus (mSTG/S) is specifically sensitive to highly-familiar, overlearned, speech categories, responding more strongly to native vowels and syllables relative to spectrotemporally matched non-speech sounds (Liebenthal et al., 2005; Joanisse et al., 2007; Obleser et al., 2007; Leaver and Rauschecker, 2010; Turkeltaub and Coslett, 2010; DeWitt and Rauschecker, 2012), or relative to non-native speech sounds (Jacquemot et al., 2003; Golestani and Zatorre, 2004). Importantly, there appears to be spatial segregation within the left STG, such that dorsal STG areas largely surrounding the auditory core demonstrate sensitivity to acoustic features relevant to phonetic perception (whether embedded within speech or non-speech sounds), and a comparatively small ventral STG area adjoining the upper bank of the middle superior

temporal sulcus (mSTG/S) demonstrates specificity to phonemic processing (Humphries et al., 2013). Thus, there is evidence for hierarchical organization of a ventral stream of processing in the left superior temporal cortex for the representation of phonemic information based on acoustic phonetic features.

These findings indicate at least two levels of processing for auditory phonemic information in the left lateral STG, generally consistent with the hierarchical processing of spectral and temporal sound structure during auditory object perception in belt and parabelt areas in the monkey (Rauschecker, 1998; Kaas and Hackett, 2000; Rauschecker and Tian, 2000; Rauschecker and Scott, 2009). In the monkey, selectivity for communication calls has been shown in the lateral belt (Rauschecker et al., 1995) and especially in the anterolateral area feeding into the ventral stream (Tian et al., 2001), already one synaptic level from the core, although it is possible that increased selectivity occurs along the ventral-stream hierarchy. In the human, it appears that selectivity for phoneme processing in the left mSTG/S is at least two synaptic levels downstream from the auditory core. An important implication of the foregoing findings for our discussion here is that neural representations of auditory speech features in the left STG are *modal* (and not a-modal or symbolic), as they preserve a form of the acoustic signal that is abstracted from low-level acoustic characteristics coded in hierarchically earlier auditory cortex. This intermediate level of sensory information representation (preserving the form of complex sensory features or patterns) is predicted by a computational model of categorical auditory speech perception (Harnad, 1987). The findings are also consistent with models of speech perception based primarily on acoustic features (Stevens and Wickesberg, 2002). An open question however, is how to correctly characterize neural representations in the phonemic left mSTG/S area. The anatomical proximity of this area to auditory cortex and strong specificity for speech perception over other language functions (Liebenthal et al., 2014) may suggest retention of some acoustic form (though greatly abstracted) even at this higher level of the speech processing hierarchy. Activation in areas more anterior in the STG (relative to mSTG/S) has been associated with the processing of linguistic and paralinguistic features available in larger chunks of speech such as words and sentences, for example syntax, prosody, and voice (Belin et al., 2000; Zatorre et al., 2004; Humphries et al., 2005, 2006; Hoekert et al., 2008; DeWitt and Rauschecker, 2012), whereas activation in the more ventral middle temporal cortex is associated with speech comprehension (Binder, 2000; Binder et al., 2000; Scott et al., 2000; Davis and Johnsrude, 2003; Humphries et al., 2005; DeWitt and Rauschecker, 2012).

Other areas outside the left mSTG/S have also been implicated in the neural representation of auditory phonemic information, particularly during phonological processing (i.e., when phonemic perception involves phonological awareness and phonological working memory, for example during explicit phonemic category judgment). The areas implicated in phonological processing are primarily those associated with the auditory dorsal pathway, including the posterior superior temporal gyrus (pSTG), inferior parietal cortex and ventral aspect of the precentral gyrus (Wise et al., 2001; Davis and Johnsrude, 2003; Buchsbaum et al., 2005; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Liebenthal et al., 2010, 2013). Neurons in the supramarginal gyrus (SMG) (Caplan et al., 1997; Celsis et al., 1999; Jacquemot et al., 2003; Guenther et al., 2006; Raizada and Poldrack, 2007; Desai et al., 2008; Tourville et al., 2008) and ventral precentral gyrus (Wilson and Iacoboni, 2006; Meister et al., 2007; Chang et al., 2010; Osnes et al., 2011; Chevillet et al., 2013) may represent the somatosensory and motor properties of speech sounds, and these areas are thought to exert modulatory influences on phonemic processing. In the inferior frontal cortex (pars opercularis in particular), sensitivity to phoneme categories (Myers et al., 2009; Lee et al., 2012; Niziolek and Guenther, 2013) may be related to the role of more anterior inferior frontal cortex areas (pars orbitalis, pars triangularis) in response selection during auditory and phoneme categorization tasks.

The evidence reviewed here is consistent with the idea that both ventral and dorsal auditory streams contribute to phonemic perception. Phonemic perception in the left ventral auditory stream is organized hierarchically from dorsal STG areas surrounding the auditory core and representing acoustic phonetic features to ventral mSTG/S areas representing phoneme categories. In the dorsal auditory pathway, phonemic perception is a result of the interaction of neurons in the left pSTG representing acoustic phonetic features of speech and neurons in inferior parietal and frontal regions representing somatosensory and motor properties of speech. With respect to visual speech, the strategic location of pSTG at the junction with inferior parietal and ventral motor cortex and the multifunctionality of this area (Liebenthal et al., 2014) make it ideally suited to interact with visual speech areas and mediate the effects of visual speech input on auditory phonemic perception, an observation that has been extensively explored in the audiovisual speech processing literature, which we discuss below. However, visual speech may also exert its influence through interaction with frontal cortices, also discussed below.

INTERIM SUMMARY

Research on auditory speech is producing a detailed understanding of the organization of auditory speech representations. Although far from complete, the present view is that auditory speech is processed hierarchically from basic acoustic feature representations, to phonetic features and phonemes, and then to higher-levels such as words. The evidence is strong that neural representations of auditory speech features in the left STG are modal (and not a-modal or symbolic), as they preserve an acoustic form of the signal that is abstracted from low-level acoustic characteristics coded in hierarchically earlier auditory cortex. This evidence has at least one very strong implication for visual speech perception: Visual speech is not expected to share representations with auditory speech at its early modal levels of representation.

MULTISENSORY SPEECH PROCESSING RESEARCH: ITS RELEVANCE TO UNDERSTANDING VISUAL SPEECH REPRESENTATIONS

Evidence is abundant that the brain is remarkably multisensory (Foxe and Schroeder, 2005; Schroeder and Foxe, 2005; Ghazanfar and Schroeder, 2006; Kayser et al., 2012), in the sense that it affords diverse neural mechanisms for integration and/or interaction (Stein et al., 2010) among different sensory inputs.

Research on audiovisual speech processing has focused on discovering those mechanisms. But the approaches have mostly not been designed to answer questions about the organization of unisensory speech representations: It has focused on answering questions such as whether there are influences from visual speech in classically defined auditory cortical areas (e.g., Sams et al., 1991; Calvert et al., 1997, 1999; Bernstein et al., 2002; Pekkola et al., 2005), whether relative information clarity in auditory vs. visual stimuli affects neural network activations (Nath and Beauchamp, 2011; Stevenson et al., 2012), and whether audiovisual integration demonstrates the principle of inverse effectiveness [(Stein and Meredith, 1993) i.e., multisensory gain is inversely related to unisensory stimulus effectiveness] (e.g., Calvert, 2001; Beauchamp, 2005; Stevenson et al., 2012). Studies of multisensensory speech interactions commonly depend on designs that use audiovisual, auditory-only, and visual-only speech stimuli without controls designed to test hypotheses about the detailed organization of unisensory processing. Unisensory stimuli are used in the research as controls and for defining multisensory sites. For example, a common control for visual-only speech is a still frame of the talker or a no-stimulus baseline (e.g., Sekiyama et al., 2003; Stevenson and James, 2009; Nath and Beauchamp, 2011, 2012; Barros-Loscertales et al., 2013; Okada et al., 2013).

Because of the interest in multisensory interactions, research has focused on putative integration sites such as the pSTS (Calvert et al., 2000; Wright et al., 2003; Callan et al., 2004; Nath and Beauchamp, 2012; Stevenson et al., 2012), which is part of both the auditory and visual pathways (see **Figure 1**). The left pSTS is routinely activated during audiovisual phoneme perception (e.g., Calvert, 2001; Sekiyama et al., 2003; Miller and D'Esposito, 2005; Stevenson and James, 2009; Nath and Beauchamp, 2011). However, high-resolution examination of pSTS demonstrates clusters of neurons in the dorsal and ventral bank of bilateral pSTS that respond to either auditory or visual input, with intervening clusters responding most strongly to audiovisual input (Beauchamp et al., 2004). What speech pattern attributes may be coded by such multisensory vs. unisensory clusters has not to our knowledge been investigated. In monkey, the STS has been found to have stronger feedback, as well as feed forward, connections with auditory and visual association rather than core areas (Seltzer and Pandya, 1994; Lewis and Van Essen, 2000; Foxe et al., 2002; Ghazanfar et al., 2005; Smiley et al., 2007).

INTERIM SUMMARY

To this point, we have reviewed the evidence that demonstrates visual perception of every psycholinguistic level of speech stimuli. We have discussed the hypothesis that visual speech might be represented through the auditory speech pathway. But our review of the auditory speech pathways suggests that representations are considered to be modal to the level of phonetic and phonemic speech representations in superior temporal regions beyond auditory core. Our view of the audiovisual speech processing literature is that its focus on multisensory interactions has resulted in limited evidence about the organization of the unisensory speech pathways. However, the expectation from the study of pSTS is that visual speech representations are projected to pSTS, and the question then is what information is represented through the visual system.

ORGANIZATION OF THE BOTTOM-UP VISUAL PATHWAYS AND IMPLICATIONS FOR SPEECH REPRESENTATIONS

Since the 1980s, the visual system organization has been described in terms of a *ventral* stream associated with form and object perception, and a *dorsal* stream associated with movement, space perception, and visually guided actions (Ungerleider and Mishkin, 1982; Goodale et al., 1994; Ungerleider and Haxby, 1994; Logothetis and Sheinberg, 1996; Zeki, 2005). Both streams effect hierarchical organization with each level of representations building on preceding ones, and higher levels are more invariant to surface characteristics of visual objects, such as orientation and size. But perception is not limited to higher level representations. That is, perceivers have access to multiple levels of the pathways (Hochstein and Ahissar, 2002; Zeki, 2005).

In its general outline, the visual ventral stream extends from V1 in the occipital lobe to V2, V3, and V4, and into ventral temporal cortex and frontal cortex. The dorsal stream extends from V1 into V2, V3, V5/MT, and dorsal temporal areas including STS, extending further to parietal and frontal areas. This organization has long been known to be not strictly hierarchical and to comprise cross-talk among areas (Felleman and Van Essen, 1991; for a recent review, Perry and Fallah, 2014). A recent proposal for a three-stream model (Weiner and Grill-Spector, 2013) implicates communication between ventral and dorsal streams for language processing, to which we return below.

VISUAL PATHWAY ORGANIZATIONS OF FACES, ORTHOGRAPHY, AND SIGN LANGUAGE PERCEPTION

The organization of visual speech pathways could possibly be in common with the organization of other types of input, including faces, orthography, and possibly sign language that share certain attributes with visual speech. Face processing obviously must to be considered in relationship to visual speech (Campbell et al., 1986; Campbell, 2011). Faces and visual speech are usually co-present, and faces are a rich source of many types of socially significant information (Allison et al., 2000; Haxby et al., 2002)—such as person identity, emotion, affect, and gaze. The "core face processing network" is generally considered to include the right lateral portion of the fusiform gyrus (FG) referred to as the fusiform face area (FFA), the lateral surface of the inferior occipital gyrus referred to as the occipital face area (OFA), and an area of the pSTS (Kanwisher et al., 1997; Fox et al., 2009). There is ample evidence that face and body representations are distinct (Downing et al., 2006; Weiner and Grill-Spector, 2013), and that body and visual speech representations are distinct (Santi et al., 2003). Face areas in cortex may be localized more reliably with moving than with still face stimuli (Fox et al., 2009). In a comparison between static and dynamic non-speech face images, right FFA and OFA did not prefer dynamic images but right posterior and anterior STS did (Pitcher et al., 2011). However, in a study with different frame rates and scrambled vs. ordered frames of non-speech facial motion stimuli, differential effects were observed in face processing areas (Schultz et al., 2013): Bilaterally, STS was more responsive to dynamic and ordered

frames, but FFA and OFA were not sensitive to the order of frames, only to the amount of image diversity in the scrambled frames.

Visual speech activations have also been recorded in the FG (Calvert and Campbell, 2003; Capek et al., 2008), leading to the suggestion that visual speech processing uses the FFA (Campbell, 2011). However, as noted above, the moving face is likely to more effectively activate face representations in the FFA, and diverse static images activate FFA more effectively than a single image. An independent face localizer is needed to functionally define the FFA region of interest (ROI) (Kanwisher et al., 1997), because it cannot be defined based on anatomy alone. But FFA localizers have not typically been used with visual speech. To determine whether FFA represents speech distinctions such as speech features or phonemes also requires methods that are sensitive to differences across speech features or phonemes within FFA ROIs. Below, we discuss results when an independent FFA localizer was used, and FFA was shown responsive to speech stimuli but less so than to non-speech face movements (Bernstein et al., 2011).

Although orthography is visually different from visual speech, both stimulus types likely make contact with higher-level mechanisms of spoken language; and both may involve recognizing words through fairly automatized whole-word recognition and also phonological analyses. Dorsal and ventral pathways have been shown to represent orthographic stimuli (Pugh et al., 2000; Jobard et al., 2003; Borowsky et al., 2006). With respect to language, as with the auditory ventral pathway, the visual ventral pathway organized from occipital through inferior temporal to frontal regions is characterized as having responsibility for relating orthographic forms to word meanings. The ventral stream could be viewed as representing specifically the forms of familiar words and exception words (e.g., letter strings with atypical spelling-to-sound correspondences, e.g., "pint"), and mapping them to word pronunciations.

We are not suggesting that lipreading is built on reading. If anything, the opposite would be more likely, given that speech is encountered earlier in development, and given that orthography is an evolutionarily recent form of visual input. But the dual stream organization observed in reading research could be related to the processing resources needed by lipreaders, inasmuch as a more skilled lipreader would be expected to have more automatized access to certain lexical items as well as need for phonological processing; and a less skilled lipreader might have greater reliance on dorsal stream processing to glean fragmentary phonetic or phonemic category information and construct possible lexical items in stimuli. Spoken words with few or no visually similar competitors (Auer and Bernstein, 1997; Iverson et al., 1998) might be particularly good candidates for skilled lipreading via whole-word representations. Likewise, the wide individual differences among lipreaders (Bernstein et al., 2000; Auer and Bernstein, 2007) could be the consequence of differential development of visual speech pathways.

Sign language perception is also visually distinct from visual speech but might have some commonality with lipreading. Classical language areas (inferior frontal and posterior temporal areas) within the left hemisphere were recruited by American Sign Language in deaf and hearing native signers (Bavelier et al., 1998). However, lipreading, auditory speech perception, and reading are united by their basis in spoken language (MacSweeney et al., 2008). In addition, deaf users of sign language likely have experienced extensive neuroplastic changes in cortical and sub-cortical organization (MacSweeney et al., 2004; Fine et al., 2005; Auer et al., 2007; Kral and Eggermont, 2007; Lyness et al., 2014) such that there could be commonality in the visual pathway for representing the configurations and dynamics of visual speech and signs. Both types of stimuli are reliant on form and motion. But research on sign language processing emphasizes commonalities at higher psycholinguistic levels (MacSweeney et al., 2002). However, consistent with reading, there is some evidence for dual-stream processing of sign language. Hearing native signers activated left inferior termporal gyrus (ITG) and STS more with British sign language than with Tic Tac, a manual system used by bookmakers at race tracks (MacSweeney et al., 2004) in contrast with hearing non-signers. Hearing native signers more than non-native signers activated ITG and middle temporal gyrus (MTG) for word lists vs. a still baseline, supporting a general role for the ventral pathway in fluent word recognition regardless of the form of the stimuli (speech, sign, orthography).

ORGANIZATION OF VISUAL SPEECH PROCESSING

In our model of auditory and visual modality-specific processing (**Figure 1**), we assume the standard visual pathways labeled "dorsal" and "ventral," because we expect that visual speech is subject to visual system organization. But the pathway labeled "dorsal" may actually correspond to the lateral pathway in Weiner and Grill-Spector (2013), which we discuss further below. The model is highly schematized, because in fact there are few results in the literature that speak directly to how the levels of speech that can be perceived by vision are neurally represented.

The literature on visual speech processing is fairly consistent in showing bilateral posterior activation in areas associated with ventral and dorsal visual pathways (Calvert et al., 1997; Campbell et al., 2001; Nishitani and Hari, 2002; Skipper et al., 2005; Bernstein et al., 2008a, 2011; Capek et al., 2008; Murase et al., 2008; Okada and Hickok, 2009; Ponton et al., 2009; Files et al., 2013). When spoken digits were contrasted with gurning (Campbell et al., 2001), bilateral FG, and right STG and MTG were more activated by speech; left IT areas were more active in the contrast between speech and a still face. When still images of speech gestures were contrasted against the baseline of a still face, bilateral FG, occipito-temporal junction, MTG, and left STS were activated (Calvert and Campbell, 2003); and dynamic stimuli were more effective than still speech in those same areas, except the bilateral lingual gyri. In a study in which spoken words were contrasted with a still face image (Capek et al., 2008), widespread bilateral activation was reported in ventral and lateral temporal areas. In a magnetoencephalography study (Nishitani and Hari, 2002), still speech images evoked a progression of activation from occipital to lateral temporal cortex labeled as pSTS. In a study in which short sentences were contrasted with videos of gurning and also with static faces (Hall et al., 2005), there was extensive bilateral but greater left-hemisphere activation in ventral and lateral middle temporal cortices. MTG activation extended to the pSTS. When lipreading syllables and gurning were contrasted (Okada

and Hickok, 2009), left posterior MTG/STS, and STG activation was obtained. When participants were imaged with positron emission tomography (PET) (Paulesu et al., 2003) while watching a still face, a face saying words, and the backwards video of the same words (backwards and forwards speech contains segments that are not different, such as vowels and transitions into and out of consonants), activations were obtained bilaterally in STG, bilateral superior temporal cortex and V5/MT. Connected speech in a story was presented in a lipreading condition that did not require any attempt to understand the story (Skipper et al., 2005), however significant activity was restricted to occipital gyri and right ITG. This result seems difficult to interpret in light of the possibility that participants were not paying attention to the speech information.

Several generalizations can be made about the above studies. A variety of stimuli was contrasted mostly against a fixed image or gurning. For the most part, visual speech stimuli reliably activated areas that can be identified within the classical ventral and dorsal visual streams. Activity was typically widespread. Activations were often bilateral although not in strictly homologous locations. Typically, results were reported as group averages and smoothed activations. Cortical surface renderings of individual activations on native anatomy were not presented. So the published results are not very helpful with regard to individual differences in anatomical location or extent of activation. Independent functional localizers for visual areas such as the FFA and V5/MT were not used, although activations generally consistent with their locations were discussed. As a group, these studies provide confirmation that the ventral and dorsal visual pathways can be activated by visual speech, but they were not designed to investigate in any detail how visual speech is represented through the pathways. To do so would have required using various controls for low-level features and higher-level objects such as faces, taking into account factors such as sensitivity to movement in FFA, using contrasts reflective of the organization of speech such as between different phonemes or speech levels, and taking into account individual variations in visual speech perception.

Bernstein et al. (2011) sought to begin to address several of the previous limitations in methodology that limit ability to determine the organization of visual speech representations in high-level vision. They used functional localizers, a variety of speech, non-speech, and moving control stimuli, and contrasted video vs. point-light images. Participants underwent independent localizer scans for the FFA, the lateral occipital complex (LOC) associated with image structure (Grill-Spector et al., 2001), and the V5/MT motion processing areas. The experimental stimuli were nonsense syllables that were selected for their visual dissimilarity ["du," "sha," "zi," "fa," "ta," "bi," "wi," "dhu" (i.e., the voiced "th"), "ku," "li," and "mu"]. In separate conditions, a variety of non-speech face gestures ("puff," "kiss," "raspberry," "growl," "yawn," "smirk," "fishface," "chew," "gurn," "nose wiggle," and "frown-to-smile") was presented. A parallel set of stimuli and controls was created based on 3-dimensional optical recordings that were made simultaneously with the video recordings. The optical recordings were of the motion of retro-reflectors positioned at 17 locations with most positions around the mouth, jaw, and cheeks. The optical recordings were used to generate point-light videos (Johansson, 1973). The point-light stimuli presented speech and non-speech motion patterns without other natural visual features such as the talker's eye gaze, shape of face components (mouth, etc.) and general appearance. Speech and non-speech stimuli were easy to discern in the point-light displays. The point-light stimulus patterns were hypothesized to represent the structure of the speech information in motion and to some extent also configuration in terms of the arrangement of the dots and shape from motion (Johansson, 1973). Point-light speech stimuli enhance the intelligibility of acoustic speech in noise (Rosenblum et al., 1996) and can interfere with audiovisual speech perception when they are incongruent (Rosenblum and Saldana, 1996). Visual controls were created from the speech and non-speech stimuli by dividing the area of the mouth and jaw into 100 square tiles. The order of frames within each tile was scrambled across sequential temporal groups of three frames. Using this scheme, the stimulus energy/luminance of the original stimuli was maintained. The control stimuli had the appearance of a face with square patches of unrelated movement.

The results showed that *non-speech* face gestures significantly activated the FFA, LOC, and V5/MT ROIs more strongly than *speech* face-gestures, supporting the expectation that none of those visual areas are selective for speech patterns. Detailed analysis of the motion data from the optical image recordings suggested that the reduced activity to speech in FFA, LOC, and V5/MT ROIs was not due to different speed of motion across stimulus types. One surprise, given its ubiquity in the literature, was that the gurn stimulus had much higher motion speed than the speech or the other non-speech stimuli. However, removal of the results that were obtained when gurns were presented did not change the overall pattern of results in ROIs.

The main experimental results were used to search for areas selective for speech independent of media (that is across point-light and video stimuli). Because point-light stimuli present primarily motion information with very much reduced configural information and no face detail, activations in conjunctions were interpreted as areas most concerned with speech patterns. Although there were activations in the right temporal cortices, the left-hemisphere activations were viewed as candidates for visual speech representations in high-level vision areas feeding forward into left-lateralized language areas. Based on individual and group results, contiguous areas of posterior MTG and STS were shown to be selective for speech. The localized posterior temporal speech selective area was dubbed the temporal visual speech area (TVSA). **Figure 1** shows the approximate location of TVSA, with the caveat that precise locations varied with individual anatomy (see Supplementary Figure 7, Bernstein et al., 2011, for individual ROIs). On an individual-participant basis, the speech activations in pSTS/pMTG were more anterior than adjacent cortex that preferred non-speech gestures. They demonstrated preliminary evidence for a positive correlation with individual lipreading scores. The finding of a visual speech area (i.e., TVSA) posterior and inferior to pSTS is consistent with the idea that TVSA is a modal area in high-level vision, possibly distinct from multisensory pSTS.

In order to examine sensitivity to phonemic speech dissimilarity in the putative TVSA, Files et al. (2013) used a visual

mismatch negativity (vMMN) paradigm to present consonant-vowel stimuli. The vMMN is elicited by change in the regularity of a sequence of visual stimuli (Pazo-Alvarez et al., 2003; Winkler and Czigler, 2012). Visual speech stimuli were selected to be *near* (ambiguous yet phonemically discriminable) or *far* (clearly different phonemes) in physical and speech perceptual distance based on a quantitative model of visual speech dissimilarity (Jiang et al., 2007). The hypothesis was tested that the left posterior temporal cortex (i.e., TVSA) has tuning for visual speech, but the right homologous cortex has tuning for discriminable speech stimuli regardless of whether they can be labeled reliably as different phonemes. Discrimination among speech stimuli that are phonemically ambiguous would be expected of cortical areas that process non-speech face movements that can vary continuously (Puce et al., 2000, 2003; Miki et al., 2004; Thompson et al., 2007; Bernstein et al., 2011) such as with different extent of mouth opening or with different motion velocities. The prediction was that regardless of perceptual distance the right hemisphere would generate the vMMN across discriminable stimuli; but only *far* phonemic contrasts would generate the vMMN on the left. Larger, more discriminable phoneme differences would be expected to feed forward to the left-lateralized language cortex.

Several attempts had previously been made to obtain vMMNs for visual speech category differences (Sams et al., 1991; Colin et al., 2002, 2004; Saint-Amour et al., 2007; Ponton et al., 2009; Winkler and Czigler, 2012). In those studies, either the vMMN was not obtained, the mismatch response was at a very long latency suggesting that it was not related to input pattern processing *per se*, or the obtained vMMN could be attributed to non-speech visual stimulus attributes. In Files et al. (2013), the stimulus selection was designed to defend against mismatch responses due to stimulus differences other than phoneme membership (be it perceptually near or far). Two tokens were presented for each phoneme category so that the vMMN would not be attributable to individual stimulus token differences. Stimuli were shifted spatially from trial to trial to defend against low-level stimulus change such as slight head or eye position variation on the screen. Care was taken to identify the temporal points in each stimulus at which the moving speech images deviated from each other, and those points were used to measure the vMMN latencies.

Current density reconstructions (Fuchs et al., 1999) and statistical analyses using clusters of posterior temporal electrodes showed reliable left-hemisphere responses to individual stimuli and vMMNs to *far* stimulus phonemic category change. On the right, vMMNs were obtained with both *far* and *near* changes. Responses were in the range of latencies observed with non-speech face gestures stimuli. Current density reconstructions demonstrated consistent patterns of posterior temporal responses in the region of pMTG to the visual speech stimuli (Figures 4–6 in Files et al., 2013), with the caveat that reconstructions are limited in their spatial resolution. The finding of hemispheric differences in the pattern of vMMN responses, with greater sensitivity to smaller difference on the right, was interpreted as evidence the left posterior temporal cortex (putative TVSA) processes phonemic patterns that feed forward into language processing areas, and that more analog processing is carried out on the right as would be required for perceiving non-categorical, non-speech face gestures.

PROPOSED MODEL

Figure 1 proposes a schematic model of the auditory and visual pathways and interactions between them. The primary prediction of the model is that modal representations of visual speech exist to the level of the TVSA, and that this area is posterior and ventral to the multisensory pSTS. We acknowledge that far too little experimental evidence currently exists to determine with any precision what the organization of visual speech representations is through the visual system.

Lipreading must rely on processing of both configural features and/or stimulus patterns, and dynamic stimulus features. Although the processing of configural features is typically associated with the ventral visual stream and that of dynamic features with the dorsal visual stream, both types of information may be represented along both ventral and dorsal streams to some extent. Form has long been known to be perceived from motion (Johansson, 1973). Current research on interactions between dorsal and ventral stream processing in object and motion perception (for a review see Perry and Fallah, 2014) supports the view that object segmentation and representation is assisted by motion features, and motion representations are affected by object form input. Perry and Fallah propose that these interactions may occur further downstream from the visual motion area (MT). The conjunction results in Bernstein et al. (2011) using point-light and video speech stimuli that localized TVSA in pMTG seems consistent with the suggestion that TVSA is responsive to both form and motion. Observations of speech activations in IT could be due to configural processing but likely are supported by motion processing, given cross-talk between ventral and dorsal streams.

It is an entirely open question whether the identified TVSA has an internal organization that could support processing in both the dorsal and ventral visual streams, for example, as an anterior area that is part of the ventral stream and a posterior area that is part of the dorsal stream, similar to the anterior-to-posterior differentiation in the left STG for auditory speech perception. It also remains an open question whether TVSA overlaps at least partially with other high-level visual areas, for example LOC in the ventral visual stream. We suggest that such questions can be answered only with careful mapping of the different functional areas within individuals and taking into account perceptual variability.

Recently, a three-stream model was proposed by Weiner and Grill-Spector (2013). In their model, the visual system is organized in terms of a dorsal vision-action stream, a ventral visual perception stream for recognition of forms such as objects and faces, and a lateral stream concerned with form, visual dynamics and language, among other functions. The lateral pathway comprises the lateral occipital sulcus, the middle occipital gyrus, the posterior inferior temporal sulcus, and the MTG extending into V5/MT. The lateral stream communicates with both the parietal cortex of the dorsal stream and the inferior temporal cortex of the ventral stream. This arrangement is compatible with what is known to date about visual speech processing. Weiner and Grill-Spector do not elaborate on the possible role of their proposed lateral stream, but research on visual speech processing

could contribute to a better understanding of this proposed lateral pathway.

THE ROLE OF FRONTAL AND PARIETAL AREAS IN VISUAL SPEECH PERCEPTION

Our discussion of a neural model of visual speech perception has focused thus far on high-level vision areas. However, as for auditory speech perception, other motor and somatosensory areas in the frontal and parietal cortex have also been implicated in visual speech perception, particularly within the theoretical framework that posits a human frontal cortex mirror neuron system (Rizzolatti and Arbib, 1998). This view is compatible with the longstanding motor theory of speech perception (Liberman and Mattingly, 1985) and with the evidence for modulatory effects of the somatomotor system on auditory phonemic perception reviewed above (Wilson et al., 2004; Meister et al., 2007; Möttönen and Watkins, 2009; Osnes et al., 2011) in the context of a somatomotor role for both the auditory and visual dorsal streams (Rauschecker and Scott, 2009).

Frontal cortex activation is commonly observed with audiovisual or visual speech perception (e.g., MacSweeney et al., 2000; Bernstein et al., 2002, 2011; Möttönen et al., 2002; Callan et al., 2003; Calvert and Campbell, 2003; Paulesu et al., 2003; Sekiyama et al., 2003; Miller and D'Esposito, 2005; Ojanen et al., 2005; Skipper et al., 2005, 2007b; Okada and Hickok, 2009; Matchin et al., 2014). Inferior frontal activations during overt categorization of speech stimuli have been attributed to a role of this area in cognitive control and domain-general category computation (Hasson et al., 2007; Myers et al., 2009). Somatomotor system engagement is often observed in the context of failure to integrate audiovisual stimuli. Because visual speech is typically less intelligible than acoustic speech, or is presented in the context of noisy acoustic speech, speech somatomotor activity observed during audiovisual speech perception could arise due to conflict resolution with degraded speech (Miller and D'Esposito, 2005; Callan et al., 2014) or due to response biases (Venezia et al., 2012). However, unlike auditory and visual cortices, the frontal cortex does not appear to play a critical role in the perception of clear speech, that is, in the accurate representation of stimulus patterns.

A study (Hasson et al., 2007) comparing rapid adaptation (Grill-Spector and Malach, 2001) effects with veridical vs. perceptual speech stimulus repetition concluded that areas in inferior frontal gyrus (IFG) coded for perceptual rather than sensory physical stimulus properties. Thus, when a mismatched visual "ka" and auditory "pa" were preceded by an audiovisual "ta"—the syllable typically heard with the mismatched stimuli—adaptation in IFG was similar to that with a veridical audiovisual "ta." Thus, the observed adaptation effects followed perceived category change and not sensory stimulus change.

Callan et al. (2014) presented CVC English words under audiovisual conditions with three levels of noise, auditory-only conditions with three levels of noise, visual-only speech, and a still face baseline. The task was forced-choice identification of the vowel. Visual-only and audiovisual stimuli activated left IFG and ventral premotor cortex. Visual-only activation was greater than audiovisual in a dorsal part of the premotor cortex, implying some modal effects even in frontal cortex. However, there was not an examination of categorization effects within the dorsal premotor cortex, so it is not at all clear what the modality-specific response is attributable to.

The SMG has also been a focus in research on audiovisual speech integration (Hasson et al., 2007; Bernstein et al., 2008a,b; Arnal et al., 2009; Dick et al., 2010). Activation in this area has been observed with visual-only speech (Chu et al., 2013) and with auditory speech (Caplan et al., 1997; Celsis et al., 1999; Jacquemot et al., 2003; Guenther et al., 2006; Raizada and Poldrack, 2007; Desai et al., 2008; Tourville et al., 2008; Liebenthal et al., 2013). Left SMG is sensitive to individual differences in processing incongruity of visual speech (Hasson et al., 2007). It is sensitive to the degree of stimulus incongruity measured independently across auditory and visual speech, which suggests also that some modal aspect of representation extends to the SMG (Bernstein et al., 2008b).

Overall, common activation in parietal and frontal areas in response to auditory and visual speech is expected (see **Figure 1**), in light of the evidence that such areas participate in higher-level (amodal) aspects of language processing.

SUMMARY AND CONCLUSIONS

Our inquiry into the visual speech perception literature shows that all levels of speech patterns that can be heard can also be seen, with the proviso that perception is subject to large individual differences. The perceptual evidence is highly valuable, because it leads to a strong rationale for undertaking research to discover how the brain represents visual speech.

We discussed the implication from neuroimaging results that visual speech has special status in possibly being represented not by the visual system but by the auditory system. Our review of the literature, including the organization of the auditory pathways leads us to doubt the validity of that suggestion. Modal representations of auditory speech exist beyond the auditory core areas that have been observed to respond to visual speech. We are in accord with the view that those activations are related to feedback, modulatory effects (Calvert et al., 1999) and not to the representation of visual speech patterns *per se*.

Neuroimaging literature on lipreading shows widespread and diverse activity in the classical ventral and dorsal visual pathways in response to visual speech. However, the literature has for the most part not addressed in sufficient detail the organization and specificity of visual pathways for visual speech perception. A main drawback has been the use of baseline stimuli such as a still face or gurns to contrast with visual speech. Our recent fMRI and EEG studies with more in-depth focus on visual speech attributes provide evidence for a left posterior temporal area, TVSA, in high-level vision, possibly the recipient of both ventral and dorsal stream input, and sensitive to phonetic and phonemic speech attributes.

While there is not at the moment sufficient evidence for making detailed neuroanatomical predictions regarding the organization of the visual cortex for visual speech processing, we make the following empirically testable predictions: (1) The visual perception of speech relies on visual pathway representations of speech *qua* speech. That is, visual speech perception relies on stimulus patterns represented through visual pathways. (2) A proposed

site of these, the TVSA, has been demonstrated in posterior temporal cortex, ventral and posterior to multisensory posterior superior temporal sulcus (pSTS). TVSA may feed modal information to downstream multisensory integration sites in pSTS. (3) Given that visual speech has dynamic and configural features that together are important for visual speech perception, neural representation of visual speech in feed forward visual pathways are expected to integrate to some extent across these features, possibly at the level of TVSA. Thus, a rigid division of the visual system into a dorsal and a ventral stream likely is not an adequate description for visual speech. Rather, the expectation is that there is cross-talk between areas in these paths for the processing of visual speech. (4) Visual speech information is expected to be fed forward from the occipital cortex to both the inferior parietal cortex along a dorsal visual pathway, and to the middle temporal cortex along a ventral visual pathway. Given the implication of the occipital-parietal (dorsal) visual stream in visual control of motor actions and spatial short-term memory (amongst other functions), we expect that the neural representations of visual speech in high-level areas of this stream may maintain more of the veridical, dynamic, and sequential information of the visual input, similar to neural representations of speech in the dorsal auditory stream (Wise et al., 2001; Buchsbaum et al., 2005; Hickok and Poeppel, 2007; Rauschecker and Scott, 2009; Liebenthal et al., 2010). Given the implication of the occipito-temporal (ventral) visual stream in visual object recognition and long-term memory, we expect that neural representations in high-level areas of this stream may be highly abstracted from the visual input, similar to the neural representations of speech phonemes in the ventral auditory pathway (Liebenthal et al., 2005; Joanisse et al., 2007; Obleser et al., 2007; Leaver and Rauschecker, 2010; Turkeltaub and Coslett, 2010; DeWitt and Rauschecker, 2012).

We make the following suggestions for future research: (1) Given individual differences in perception and functional location of TVSA, detailed examination is needed within individuals to understand the organization of visual speech representations; (2) To understand fully how neural processes underlying visual and auditory speech perception interact, examination is needed, again within individuals, of the organization of both visual and auditory pathways for speech perception. (3) The ability to visually perceive all the psycholinguistic levels of speech calls for research both within and across psycholinguistic levels (i.e., phonetic features, phonemes, syllables, words, and prosody) of organization. In principle, the organization of visual speech processing cannot be determined based only on unspecific contrasts such as speech stimuli vs. still face images.

ACKNOWLEDGMENTS

We thank the reviewers and editor for their insightful comments. This paper was supported in part by grants from the US National Institutes of Health/National Institute on Deafness and Other Communication Disorders grants DC008583, DC008308 (Bernstein PI) and DC006287 (Liebenthal, PI).

REFERENCES

Allison, T., Puce, A., and McCarthy, G. (2000). The neurobiology of social cognition. *Trends Cogn. Sci. (Regul. Ed).* 4, 267–279. doi: 10.1016/S1364-6613(00)01501-1

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

Auer, E. T. Jr. (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

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. Acous. Soc. Am.* 102, 3704–3710. doi: 10.1121/1.420402

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

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

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

Bavelier, D., Corina, D., Jezzard, P., Clark, V., Karni, A., Lalwani, A., et al. (1998). Hemispheric specialization for English and ASL: left invariance-right variability. *Neuroreport* 9, 1537–1542. doi: 10.1097/00001756-199805110-00054

Beauchamp, M. S. (2005). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics* 3, 93–113. doi: 10.1385/NI:3:2:093

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

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–312. doi: 10.1038/35002078

Bernstein, L. E. (2012). "Visual speech perception," in *AudioVisual Speech Processing*, eds E. Vatikiotis-Bateson, G. Bailly, and P. Perrier (Cambridge: Cambridge University), 21–39.

Bernstein, L. E., Auer, E. T. Jr., Moore, J. K., Ponton, C. W., Don, M., and Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *Neuroreport* 13, 311–315. doi: 10.1097/00001756-200203040-00013

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

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

Bernstein, L. E., Demorest, M. E., and Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: phoneme-to-phoneme stimulus-response alignment. *J. Acous. Soc. Am.* 95, 3617–3622. doi: 10.1121/1.409930

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

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

Bernstein, L. E., Eberhardt, S. P., and Demorest, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *J. Acous. Soc. Am.* 85, 397–405. doi: 10.1121/1.397690

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

Bernstein, L. E., Lu, Z. L., and Jiang, J. (2008b). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Res.* 1242, 172–184. doi: 10.1016/j.brainres.2008.04.018

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

Binder, J. R. (2000). The new neuroanatomy of speech perception. *Brain* 123(Pt 12), 2371–2372. doi: 10.1093/brain/123.12.2371

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

Borowsky, R., Cummine, J., Owen, W. J., Friesen, C. K., Shih, F., and Sarty, G. E. (2006). FMRI of ventral and dorsal processing streams in basic reading processes: insular sensitivity to phonology. *Brain Topogr.* 18, 233–239. doi: 10.1007/s10548-006-0001-2

Bottari, D., Heimler, B., Caclin, A., Dalmolin, A., Giard, M. H., and Pavani, F. (2014). Visual change detection recruits auditory cortices in early deafness. *Neuroimage* 94, 172–184. doi: 10.1016/j.neuroimage.2014.02.031

Buchsbaum, B. R., Olsen, R. K., Koch, P., and Berman, K. F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron* 48, 687–697. doi: 10.1016/j.neuron.2005.09.029

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

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., and Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport* 14, 2213–2218. doi: 10.1097/00001756-200312020-00016

Callan, D. 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. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

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

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

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

Campbell, R. (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. (2011). Speechreading and the Bruce-Young model of face recognition: early findings and recent developments. *Br. J. Psychol.* 102, 704–710. doi: 10.1111/j.2044-8295.2011.02021.x

Campbell, R., Landis, T., and Regard, M. (1986). Face recognition and lipreading. A neurological dissociation. *Brain* 109(Pt 3), 509–521. doi: 10.1093/brain/109.3.509

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cogn. Brain Res.* 12, 233–243. doi: 10.1016/S0926-6410(01)00054-4

Capek, C. M., Macsweeney, M., Woll, B., Waters, D., McGuire, P. K., David, A. S., et al. (2008). Cortical circuits for silent speechreading in deaf and hearing people. *Neuropsychologia* 46, 1233–1241. doi: 10.1016/j.neuropsychologia.2007.11.026

Caplan, D., Waters, G. S., and Hildebrandt, N. (1997). Determinants of sentence comprehension in aphasic patients in sentence-picture matching tasks. *J. Speech Lang. Hear. Res.* 40, 542–555. doi: 10.1044/jslhr.4003.542

Catford, J. C. (1977). *Fundamental Problems in Phonetics*. Bloomington, IN: Indiana University.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J. P., Berry, I., Nespoulous, J. L., et al. (1999). Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *Neuroimage* 9, 135–144. doi: 10.1006/nimg.1998.0389

Chan, A. M., Dykstra, A. R., Jayaram, V., Leonard, M. K., Travis, K. E., Gygi, B., et al. (2014). Speech-specific tuning of neurons in human superior temporal gyrus. *Cereb Cortex*. 24, 2679–2693. doi: 10.1093/cercor/bht127

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

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

Chu, Y.-H., Lin, F.-H., Chou, Y.-J., Tsai, K. W.-K., Kuo, W.-J., and Jaaskelainen, L. P. (2013). Effective cerebral connectivity during silent speech reading revealed by functional magnetic resonance imaging. *PLoS ONE* 8:e80265. doi: 10.1371/journal.pone.0080265

Colin, C., Radeau, M., Soquet, A., and Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. *Clin. Neurophysiol.* 115, 1989–2000. doi: 10.1016/j.clinph.2004.03.027

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

Conklin, E. S. (1917). A method for the determination of relative skill in lip-reading. *Volta Rev.* 19, 216–219.

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

Demorest, M. E., and Bernstein, L. E. (1997). Relationships between subjective ratings and objective measures of performance in speechreading sentences. *J. Speech Lang. Hear. Res.* 40, 900–911. doi: 10.1044/jslhr.4004.900

Desai, R., Liebenthal, E., Possing, E. T., Waldron, E., and Binder, J. R. (2005). Volumetric vs. surface-based alignment for localization of auditory cortex activation. *Neuroimage* 26, 1019–1029. doi: 10.1016/j.neuroimage.2005.03.024

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

DeWitt, 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

Dick, A. S., Solodkin, A., and Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain Lang.* 114, 101–114. doi: 10.1016/j.bandl.2009.08.005

Downing, P. E., Chan, A. W., Peelen, M. V., Dodds, C. M., and Kanwisher, N. (2006). Domain specificity in visual cortex. *Cereb. Cortex* 16, 1453–1461. doi: 10.1093/cercor/bhj086

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *J. Speech Hear. Res.* 14, 496–512. doi: 10.1044/jshr.1403.496

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *J. Neurosci.* 22, 5749–5759.

Falchier, A., Schroeder, C. E., Hackett, T. A., Lakatos, P., Nascimento-Silva, S., Ulbert, I., et al. (2010). Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey. *Cereb. Cortex* 20, 1529–1538. doi: 10.1093/cercor/bhp213

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Files, B. T., Auer, E. T. Jr., and Bernstein, L. E. (2013). The visual mismatch negativity elicited with visual speech stimuli. *Front. Hum. Neurosci.* 7:371. doi: 10.3389/fnhum.2013.00371

Fine, I., Finney, E. M., Boynton, G. M., and Dobkins, K. R. (2005). Comparing the effects of auditory deprivation and sign language within the auditory and visual cortex. *J. Cogn. Neurosci.* 17, 1621–1637. doi: 10.1162/089892905774597173

Fisher, C. G. (1968). Confusions among visually perceived consonants. *J. Speech Hear. Res.* 11, 796–804. doi: 10.1044/jshr.1104.796

Fisher, C. G. (1969). The visibility of terminal pitch contour. *J. Speech Hear. Res.* 12, 379–382. doi: 10.1044/jshr.1202.379

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322, 970–973. doi: 10.1126/science.1164318

Fox, C. J., Iaria, G., and Barton, J. J. (2009). Defining the face processing network: optimization of the functional localizer in fMRI. *Hum. Brain Mapp.* 30, 1637–1651. doi: 10.1002/hbm.20630

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

Foxe, J. J., Wylie, G. R., Martinez, A. S., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., et al. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *J. Neurophysiol.* 88, 540–543.

Fuchs, M., Wagner, M., Köhler, T., and Wischmann, H. A. (1999). Linear and nonlinear current density reconstructions. *J. Clin. Neurophysiol.* 16, 267–295. doi: 10.1097/00004691-199905000-00006

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., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci. (Regul. Ed).* 10, 278–285. doi: 10.1016/j.tics.2006.04.008

Golestani, N., and Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Goodale, M. A., Meenan, J. P., Bulthoff, H. H., Nicolle, D. A., Murphy, K. J., and Racicot, C. I. (1994). Separate neural pathways for the visual analysis of object shape in perception and prehension. *Curr. Biol.* 4, 604–610. doi: 10.1016/S0960-9822(00)00132-9

Green, K. P., and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Percept. Psychophys.* 45, 34–42. doi: 10.3758/BF03208030

Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Res.* 41, 1409–1422. doi: 10.1016/S0042-6989(01)00073-6

Grill-Spector, K., and Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol.* 107, 293–321. doi: 10.1016/S0001-6918(01)00019-1

Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301. doi: 10.1016/j.bandl.2005.06.001

Hall, D. A., Fussell, C., and Summerfield, A. Q. (2005). Reading fluent speech from talking faces: typical brain networks and individual differences. *J. Cogn. Neurosci.* 17, 939–953. doi: 10.1162/0898929054021175

Harnad, S. (1987). "Category induction and representation," in *Categorical Perception: The Groundwork of Cognition,* ed S. Harnad (New York, NY: Cambridge University Press), 535–565.

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

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biol. Psychiatry* 51, 59–67. doi: 10.1016/S0006-3223(01)01330-0

Haxby, J. V., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietrini, P., and Grady, C. L. (1994). The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. *J. Neurosci.* 14(11 Pt 1), 6336–6353.

Hertz, U., and Amedi, A. (2014). Flexibility and stability in sensory processing revealed using visual-to-auditory sensory substitution. *Cereb. Cortex*. doi: 10.1093/cercor/bhu010. [Epub ahead of print].

Hickok, G., Buchsbaum, B., Humphries, C., and Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *J. Cogn. Neurosci.* 15, 673–682. doi: 10.1162/089892903322307393

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

Hochstein, S., and Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 791–804. doi: 10.1016/S0896-6273(02)01091-7

Hoekert, M., Bais, L., Kahn, R. S., and Aleman, A. (2008). Time course of the involvement of the right anterior superior temporal gyrus and the right fronto-parietal operculum in emotional prosody perception. *PLoS ONE* 3:e2244. doi: 10.1371/journal.pone.0002244

Humphries, C., Binder, J. R., Medler, D. A., and Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *J. Cogn. Neurosci.* 18, 665–679. doi: 10.1162/jocn.2006.18.4.665

Humphries, C., Love, T., Swinney, D., and Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Hum. Brain Mapp.* 26, 128–138. doi: 10.1002/hbm.20148

Humphries, C., Sabri, M., Heugel, N., Lewis, K., and Liebenthal, E. (2013). Pattern specific adaptation to speech and non-speech sounds in human auditory cortex (354.21/SS7). *Soc. Neurosci.* Abstract 354.21/SS7.

Iverson, P., Bernstein, L. E., and Auer, E. T. Jr. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Commun.* 26, 45–63. doi: 10.1016/S0167-6393(98)00049-1

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., and Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *J. Neurosci.* 23, 9541–9546.

Jeffers, J., and Barley, M. (1971). *Speechreading (Lipreading)*. Springfield, IL: Charles C. Thomas.

Jesse, A., and McQueen, J. M. (2014). Suprasegmental lexical stress cues in visual speech can guide spoken-word recognition. *Q. J. Exp. Psychol. (Hove)*. 67, 793–808. doi: 10.1080/17470218.2013.834371

Jiang, J., Alwan, A., Keating, P., Auer, E. T. Jr., and Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP J. Appl. Signal Process.* 2002, 1174–1188. doi: 10.1155/S1110865702206046

Jiang, J., Auer, E. T. Jr., Alwan, A., Keating, P. A., and Bernstein, L. E. (2007). Similarity structure in visual speech perception and optical phonetic signals. *Percept. Psychophys.* 69, 1070–1083. doi: 10.3758/BF03193945

Joanisse, M. F., Zevin, J. D., and McCandliss, B. D. (2007). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using FMRI and a short-interval habituation trial paradigm. *Cereb. Cortex* 17, 2084–2093. doi: 10.1093/cercor/bhl124

Jobard, G., Crivello, F., and Tzourio-Mazoyer, N. (2003). Evaluation of the dual route theory of reading: a metanalysis of 35 neuroimaging studies. *Neuroimage* 20, 693–712. doi: 10.1016/S1053-8119(03)00343-4

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/BF03212378

Johnson, E. K., Seidl, A., and Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS ONE* 9:e83546. doi: 10.1371/journal.pone.0083546

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799. doi: 10.1073/pnas.97.22.11793

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311.

Karns, C. M., Dow, M. W., and Neville, H. J. (2012). Altered cross-modal processing in the primary auditory cortex of congenitally deaf adults: a visual-somatosensory fMRI study with a double-flash illusion. *J. Neurosci.* 32, 9626–9638. doi: 10.1523/JNEUROSCI.6488-11.2012

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

Kayser, C., Petkov, C. I., Remedios, R., and Logothetis, N. K. (2012). "Multisensory influences on auditory processing: perspectives from fMRI and electrophysiology," in *The Neural Bases of Multisensory Processes*, eds M. M. Murray and M. T. Wallace (Boca Raton, FL: CRC Press). http://www.ncbi.nlm.nih.gov/books/NBK92843/

Kilian-Hutten, N., Valente, G., Vroomen, J., and Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *J. Neurosci.* 31, 1715–1720. doi: 10.1523/JNEUROSCI.4572-10.2011

Klatt, D. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *J. Phon.* 7, 279–312.

Kral, A., and Eggermont, J. J. (2007). What's to lose and what's to learn: development under auditory deprivation, cochlear implants and limits of cortical plasticity. *Brain Res. Rev.* 56, 259–269. doi: 10.1016/j.brainresrev.2007.07.021

Kruskal, J. B., and Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage.

Kuhl, P. K., and Meltzoff, A. N. (1988). "Speech as an intermodal object of perception," in *Perceptual Development in Infancy (Vol. The Minnesota Symposia on Child Psychology, 20, pp. 235–266)*, ed A. Yonas (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.).

Lansing, C. R., and McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *J. Speech Lang. Hear. Res.* 42, 526–539. doi: 10.1044/jslhr.4203.526

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Lee, Y. S., Turkeltaub, P., Granger, R., and Raizada, R. D. (2012). Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. *J. Neurosci.* 32, 3942–3948. doi: 10.1523/JNEUROSCI.3814-11.2012

Lemus, L., Hernandez, A., Luna, R., Zainos, A., and Romo, R. (2010). Do sensory cortices process more than one sensory modality during perceptual judgments? *Neuron* 67, 335–348. doi: 10.1016/j.neuron.2010.06.015

Levanen, S., Jousmaki, V., and Hari, R. (1998). Vibration-induced auditory-cortex activation in a congenitally deaf adult. *Curr. Biol.* 8, 869–872. doi: 10.1016/S0960-9822(07)00348-X

Lewis, J. W., and Van Essen, D. C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *J. Comp. Neurol.* 428, 112–137.

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

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

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Liebenthal, E., Desai, R. H., Humphries, C., Sabri, M., and Desai, A. (2014). The functional organization of the left STS: a large scale meta-analysis of PET and fMRI studies of healthy adults. *Front. Neurosci.* 8:289. doi: 10.3389/fnins.2014.00289

Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., and Desai, A. (2013). Neural dynamics of phonological processing in the dorsal auditory stream. *J. Neurosci.* 33, 15414–15424. doi: 10.1523/JNEUROSCI.1511-13.2013

Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., and Mandler, R. (1977). On pushing the voice onset-time (VOT) boundary about. *Lang. Speech* 20, 209–216.

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Ludman, C. N., Summerfield, A. Q., Hall, D., Elliott, M., Foster, J., Hykin, J. L., et al. (2000). Lip-reading ability and patterns of cortical activation studied using fMRI. *Br. J. Audiol.* 34, 225–230. doi: 10.3109/03005364000000132

Lyness, R. C., Alvarez, I., Sereno, M. I., and MacSweeney, M. (2014). Microstructural differences in the thalamus and thalamic radiations in the congenitally deaf. *Neuroimage* 100, 347–357. doi: 10.1016/j.neuroimage.2014.05.077

Lyxell, B., Ronnberg, J., Andersson, J., and Linderoth, E. (1993). Vibrotactile support: initial effects on visual speech perception. *Scand. Audiol. Suppl.* 22, 179–183. doi: 10.3109/01050399309047465

MacSweeney, M., Amaro, E., Calvert, G. A., Campbell, R., David, A. S., McGuire, P., et al. (2000). Silent speechreading in the absence of scanner noise: an event-related fMRI study. *Neuroreport* 11, 1729–1733. doi: 10.1097/00001756-200006050-00026

MacSweeney, M., Campbell, R., Woll, B., Giampietro, V., David, A. S., McGuire, P. K., et al. (2004). Dissociating linguistic and nonlinguistic gestural communication in the brain. *Neuroimage* 22, 1605–1618. doi: 10.1016/j.neuroimage.2004.03.015

MacSweeney, M., Capek, C. M., Campbell, R., and Woll, B. (2008). The signing brain: the neurobiology of sign language. *Trends Cogn. Sci. (Regul. Ed).* 12, 432–440. doi: 10.1016/j.tics.2008.07.010

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., et al. (2002). Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain* 125, 1583–1593. doi: 10.1093/brain/awf153

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

Massaro, D. W., and Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* 9, 753–771. doi: 10.1037/0096-1523.9.5.753

Massaro, D. W., Cohen, M. M., Tabain, M., and Beskow, J. (2012). "Animated speech: research progress and applications," in *Audiovisual Speech Processing*, eds R. B. Clark, J. P. Perrier, and E. Vatikiotis-Bateson (Cambridge: Cambridge University), 246–272.

Matchin, W., Groulx, K., and Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk effect, and fMRI. *J. Cog. Neurosci.* 26, 606–620. doi: 10.1162/jocn_a_00515

Mattys, S. L., Bernstein, L. E., and Auer, E. T. Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Percept. Psychophys.* 64, 667–679. doi: 10.3758/BF03194734

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

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

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Mesulam, M. M. (1998). From sensation to cognition. *Brain* 121, 1013–1052. doi: 10.1093/brain/121.6.1013

Miki, K., Watanabe, S., Kakigi, R., and Puce, A. (2004). Magnetoencephalographic study of occipitotemporal activity elicited by viewing mouth movements. *J. Clin. Neurophysiol.* 115, 1559–1574. doi: 10.1016/j.clinph.2004.02.013

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893. doi: 10.1523/JNEUROSCI.0896-05.2005

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., and Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clin. Linguist. Phon.* 20, 621–630. doi: 10.1080/02699200500266745

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

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

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychol. Sci.* 15, 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

Murase, M., Saito, D. N., Kochiyama, T., Tanabe, H. C., Tanaka, S., Harada, T., et al. (2008). Cross-modal integration during vowel identification in audiovisual speech: a functional magnetic resonance imaging study. *Neurosci. Lett.* 434, 71–76. doi: 10.1016/j.neulet.2008.01.044

Myers, E. B., Blumstein, S. E., Walsh, E., and Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychol. Sci.* 20, 895–903. doi: 10.1111/j.1467-9280.2009.02380.x

Nath, A. R., and Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J. Neurosci.* 31, 1704–1714. doi: 10.1523/JNEUROSCI.4853-10.2011

Nath, A. R., and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787. doi: 10.1016/j.neuroimage.2011.07.024

Nishitani, N., and Hari, R. (2002). Viewing lip forms: cortical dynamics. *Neuron* 36, 1211–1220. doi: 10.1016/S0896-6273(02)01089-9

Niziolek, C. A., and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J. Neurosci.* 33, 12090–12098. doi: 10.1523/JNEUROSCI.1008-13.2013

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cereb. Cortex* 17, 2251–2257. doi: 10.1093/cercor/bhl133

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

Okada, K., and Hickok, G. (2009). Two cortical mechanisms support the integration of visual and auditory speech: a hypothesis and preliminary data. *Neurosci. Lett.* 452, 219–223. doi: 10.1016/j.neulet.2009.01.060

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., and Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PLoS ONE* 8:e68959. doi: 10.1371/journal.pone.0068959

Osnes, B., Hugdahl, K., and Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *Neuroimage* 54, 2437–2445. doi: 10.1016/j.neuroimage.2010.09.078

Owens, E., and Blazek, B. (1985). Visemes observed by hearing-impaired and normal hearing adult viewers. *J. Speech Hear. Res.* 28, 381–393. doi: 10.1044/jshr.2803.381

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., et al. (2003). A functional-anatomical model for lipreading. *J. Neurophysiol.* 90, 2005–2013. doi: 10.1152/jn.00926.2002

Pazo-Alvarez, P., Cadaveira, F., and Amenedo, E. (2003). MMN in the visual modality: a review. *Biol. Psychol.* 63, 199–236. doi: 10.1016/S0301-0511(03)00049-8

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Mottonen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport* 16, 125–128. doi: 10.1097/00001756-200502080-00010

Perry, C. J., and Fallah, M. (2014). Feature integration and object representations along the dorsal stream visual hierarchy. *Front. Comput. Neurosci.* 8:84. doi: 10.3389/fncom.2014.00084

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., and Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage* 56, 2356–2363. doi: 10.1016/j.neuroimage.2011.03.067

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

Puce, A., Smith, A., and Allison, T. (2000). ERPs evoked by viewing facial movements. *Cogn. Neuropsychol.* 17, 221–239. doi: 10.1080/026432900380580

Puce, A., Syngeniotis, A., Thompson, J. C., Abbott, D. F., Wheaton, K. J., and Castiello, U. (2003). The human temporal lobe integrates facial form and motion: evidence from fMRI and ERP studies. *Neuroimage* 19, 861–869. doi: 10.1016/S1053-8119(03)00189-7

Pugh, K. R., Mencl, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., et al. (2000). Functional neuroimaging studies of reading and reading disability (developmental dyslexia). *Ment. Retard. Dev. Disabil. Res. Rev.* 6, 207–213. doi: 10.1002/1098-2779(2000)6:3<207::aid-mrdd8>3.0.co;2-p

Raizada, R. D., and Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron* 56, 726–740. doi: 10.1016/j.neuron.2007.11.001

Raphael, L. J. (1971). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *J. Acous. Soc. Am.* 51, 1296–1303.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Curr. Opin. Neurobiol.* 8, 516–521. doi: 10.1016/S0959-4388(98)80040-8

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

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Rauschecker, J. P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111–114. doi: 10.1126/science.7701330

Risberg, A., and Lubker, J. L. (1978). "Prosody and speechreading," in *Quarterly Progress and Status Report*, Vol. 4 (Stockholm: Speech Transmission Laboratory of the Royal Institute of Technology), 1–16.

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

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

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136. doi: 10.1038/16056

Rosenblum, L. D., Johnson, J. A., and Saldana, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *J. Speech Hear. Res.* 39, 1159–1170. doi: 10.1044/jshr.3906.1159

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

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., and Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7295–7300. doi: 10.1073/pnas.0609419104

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

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

Santi, A., Servos, P., Vatikiotis-Bateson, E., Kuratate, T., and Munhall, K. (2003). Perceiving biological motion: dissociating visible speech from walking. *J. Cogn. Neurosci.* 15, 800–809. doi: 10.1162/089892903322370726

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M. S., et al. (2008). Ventral and dorsal pathways for language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18035–18040. doi: 10.1073/pnas.0805234105

Scarborough, R., Keating, P., Baroni, M., Cho, T., Mattys, S., Alwan, A., et al. (2007). *Optical Cues to the Visual Perception of Lexical and Phrasal Stress in English*. Working Papers in Phonetics, University of California, Los Angeles, CA. Available online at: http://escholarship.org/uc/item/4gk6008p.

Schroeder, C. E., and Foxe, J. J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Curr. Opin. Neurobiol.* 15, 454–458. doi: 10.1016/j.conb.2005.06.008

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

Schultz, J., Brockhaus, M., Bulthoff, H. H., and Pilz, K. S. (2013). What the human brain likes about facial motion. *Cereb. Cortex* 23, 1167–1178. doi: 10.1093/cercor/bhs106

Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123(Pt 12), 2400–2406. doi: 10.1093/brain/123.12.2400

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

Seltzer, B., and Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *J. Comp. Neurol.* 343, 445–463. doi: 10.1002/cne.903430308

Shepard, R. N., and Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. *Cogn. Psychol.* 1, 1–17. doi: 10.1016/0010-0285(70)90002-2

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2007a). Speech-associated gestures, Broca's area, and the human mirror system. *Brain Lang.* 101, 260–277. doi: 10.1016/j.bandl.2007.02.008

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

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

Smiley, J. F., Hackett, T. A., Ulbert, I., Karmas, G., Lakatos, P., Javitt, D. C., et al. (2007). Multisensory convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque monkeys. *J. Comp. Neurol.* 502, 894–923. doi: 10.1002/cne.21325

Song, J. J., Lee, H. J., Kang, H., Lee, D. S., Chang, S. O., and Oh, S. H. (2014). Effects of congruent and incongruent visual cues on speech perception and brain activity in cochlear implant users. *Brain Struct. Funct.* doi: 10.1007/s00429-013-0704-6. [Epub ahead of print].

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Meredith, A. M., Perrault, T. J., et al. (2010). Semantic confusion regarding the development of

multisensory integration: a practical solution. *Eur. J. Neurosci.* 31, 1713–1720. doi: 10.1111/j.1460-9568.2010.07206.x

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

Steinschneider, M., Nourski, K. V., Kawasaki, H., Oya, H., Brugge, J. F., and Howard, M. A. 3rd. (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cereb. Cortex* 21, 2332–2347. doi: 10.1093/cercor/bhr014

Stevens, H. E., and Wickesberg, R. E. (2002). Representation of whispered word-final stop consonants in the auditory nerve. *Hear. Res.* 173, 119–133. doi: 10.1016/S0378-5955(02)00608-1

Stevens, K. N. (1981). "Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics," in *The Cognitive Representation of Speech,* eds T. Myers, J. Laver, and J. Anderson (Amsterdam: North Holland; Elsevier Science Ltd.), 61–74.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

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

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

Strand, J. F., and Sommers, M. S. (2011). Sizing up the competition: quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *J. Acous. Soc. Am.* 130, 1663–1672. doi: 10.1121/1.3613930

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acous. 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: The Psychology of Lip-Reading,* eds B. Dodd and R. Campbell (London: Lawrence Erlbaum Associates, Inc.), 3–52.

Thompson, J. C., Hardee, J. E., Panayiotou, A., Crewther, D., and Puce, A. (2007). Common and distinct brain activation to viewing dynamic sequences of face and hand movements. *Neuroimage* 37, 966–973. doi: 10.1016/j.neuroimage.2007.05.058

Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science* 292, 290–293. doi: 10.1126/science.1058911

Tourville, J. A., Reilly, K. J., and Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage* 39, 1429–1443. doi: 10.1016/j.neuroimage.2007.09.054

Turkeltaub, P. E., and Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain Lang.* 114, 1–15. doi: 10.1016/j.bandl.2010.03.008

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., and Sommers, M. S. (2014). Lipreading in school-age children: the roles of age, hearing status, and cognitive ability. *J. Speech Lang. Hear. Res.* 57, 556–565. doi: 10.1044/2013_JSLHR-H-12-0273

Ungerleider, L. G., Courtney, S. M., and Haxby, J. V. (1998). A neural system for human visual working memory. *Proc. Natl. Acad. Sci. U.S.A.* 95, 883–890. doi: 10.1073/pnas.95.3.883

Ungerleider, L. G., and Haxby, J. V. (1994). "What" and "where" in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165. doi: 10.1016/0959-4388(94)90066-3

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems," in *Analysis of Visual Behavior,* ed D. J. Ingle (Cambridge, MA: MIT Press), 549–586.

Utley, J. (1946). A test of lip reading ability. *J. Speech Lang. Hear. Disord.* 11, 109–116. doi: 10.1044/jshd.1102.109

Van Son, N., Huiskamp, T. M. I., Bosman, A. J., and Smoorenburg, G. F. (1994). Viseme classifications of Dutch consonants and vowels. *J. Acous. Soc. Am.* 96, 1341–1355. doi: 10.1121/1.411324

Venezia, J. H., Saberi, K., Chubb, C., and Hickok, G. (2012). Response bias modulates the speech motor system during syllable discrimination. *Front. Psychol.* 3:157. doi: 10.3389/fpsyg.2012.00157

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Curr. Opin. Neurobiol.* 5, 520–526. doi: 10.1016/0959-4388(95)80014-X

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *J. Speech Hear. Res.* 20, 130–145. doi: 10.1044/jshr.2001.130

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., and Werker, J. F. (2007). Visual language discrimination in infancy. *Science* 316, 1159. doi: 10.1126/science.1137686

Weiner, K. S., and Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychol. Res.* 77, 74–97. doi: 10.1007/s00426-011-0392-x

Wilson, F. A. W., Scalaidhe, S. P. O., and Goldman-Rakic, P. S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* 260, 1955–1958. doi: 10.1126/science.8316836

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

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Winkler, I., and Czigler, I. (2012). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *Int. J. Psychophysiol.* 83, 132–143. doi: 10.1016/j.ijpsycho.2011.10.001

Wise, R. J., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., and Warburton, E. A. (2001). Separate neural subsystems within 'Wernicke's area'. *Brain* 124(Pt 1), 83–95. doi: 10.1093/brain/124.1.83

Woodward, M. F., and Barber, C. G. (1960). Phoneme perception in lipreading. *J. Speech Hear. Res.* 3, 212–222. doi: 10.1044/jshr.0303.212

Wright, T. M., Pelphrey, K. A., Allison, T., McKeowin, M. J., and McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex.* 13, 1034–1043. doi: 10.1093/cercor/13.10.1034

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Commun.* 26, 23–43. doi: 10.1016/S0167-6393(98)00048-X

Zatorre, R. J., Bouffard, M., and Belin, P. (2004). Sensitivity to auditory object features in human temporal neocortex. *J. Neurosci.* 24, 3637–3642. doi: 10.1523/JNEUROSCI.5458-03.2004

Zeki, S. (2005). The Ferrier lecture 1995: behind the seen: the functional specialization of the brain in space and time. *Philos. Trans. Biol. Sci.* 360, 1145–1183. doi: 10.1098/rstb.2005.1666

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: 25 July 2014; accepted: 10 November 2014; published online: 01 December 2014.

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

This article was submitted to the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 22 August 2014
doi: 10.3389/fnins.2014.00260

Auditory category knowledge in experts and novices

Shannon L. M. Heald*, Stephen C. Van Hedger and Howard C. Nusbaum

Department of Psychology, The University of Chicago, Chicago, IL, USA

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Peter Schneider, Heidelberg Medical School, Germany
Antoine Shahin, The Ohio State University, USA

****Correspondence:***
Shannon L. M. Heald, Department of Psychology, The University of Chicago, 5848 S. University Avenue, Chicago, IL 60637, USA
e-mail: smbowdre@uchicago.edu

What do listeners know about sounds that have a systematic organization? Research suggests that listeners store absolute pitch information as part of their representations for specific auditory experiences. It is unclear however, if such knowledge is abstracted beyond these experiences. In two studies we examined this question via a tone adjustment task in which listeners heard one of several target tones to be matched by adjusting the frequency of a subsequent starting tone. In the first experiment listeners estimated tones from one of three distributions differing in frequency range. The effect of tone matching in the three different distributions was then modeled using randomly generated data (RGD) to ascertain the degree to which individuals' estimates are affected by generalized note knowledge. Results showed that while listeners' estimates were similar to the RGD, indicating a central tendency effect reflective of the target tone distribution, listeners were more accurate than the RGD indicating that their estimates were affected by generalized note knowledge. The second experiment tested three groups of listeners who vary in the nature of their note knowledge. Specifically, absolute pitch (AP) possessors, non-AP listeners matched in musical expertise (ME), and non-AP musical novices (MN) adjusted tones from a micro-scale that included only two in-tune notes (B4 and C5). While tone estimates for all groups showed a central tendency effect reflective of the target tone distribution, each groups' estimates were more accurate than the RGD, indicating all listeners' estimates were guided by generalized note knowledge. Further, there was evidence that explicit note knowledge additionally influenced AP possessors' tone estimates, as tones closer to C5 had less error. Results indicate that everyday listeners possess generalized note knowledge that influences the perception of isolated tones and that this effect is made more evident with additional musical experience.

Keywords: categorization, expertise, audition, distributional learning

INTRODUCTION

Category knowledge is essential for making sense of our complex auditory environments. From segmenting a speech stream into meaningful units to anticipating the resolution in a musical piece, auditory categories shape our understanding and enjoyment of acoustic events.

Generally speaking, there are two broad classes of category knowledge that can be applied to auditory objects that are critical to deriving meaning from our auditory environments. The first type of conceptual knowledge, which has been referred to as an isolated concept (Goldstone, 1996), stems from a direct, associative link to an acoustic event (e.g., gun shot or dog bark). For this reason, isolated concepts are grounded in specific non-symbolic perceptual experiences (cf. Barsalou, 1993, 1999). In such cases, heard acoustic patterns may be recognized by comparison to mentally stored templates or features. For example, recognition theories that posit simple comparison of a signal against a stored prototype or exemplars of a particular category representation do so without consideration of the relationship among categories or category representations. As such, the neural instantiation of an isolated concept can be thought to be similar to the classical notion of a feature detector, whether represented as an individual cell or as population responses. Previous research has shown that there are single neurons that appear to be selective for highly complex stimuli such as faces and shapes that are object-specific although generalized over some stimulus properties (e.g., Hubel and Wiesel, 1968; Bruce et al., 1981). In these kinds of theories, the response of a feature detector need not be influenced by the states of other related feature detectors; it simply becomes activated when the trigger features are physically present. In support of this view, Freedman et al. (2003) have found evidence that neuron responses for stimulus patterns corresponding to higher-level object categories in monkey IT are feature-based and appear invariant and unaffected by the category structure to which they belong. Although, there is evidence that PFC neurons encode a variety of abstracted information including perceptual categories (Freedman et al., 2001), attentional sets (Mansouri et al., 2006), numbers (Nieder et al., 2002), and behavioral schemas (Genovesio et al., 2005).

The second class of conceptual objects is referred to as interrelated concepts (de Saussure, 1959/1916; Goldstone, 1996) and when applied to auditory objects systematically relates sound patterns to meanings in a web of knowledge, such that concepts are not solely defined in terms of their content or extensional

mapping but also in terms of their relationship with other concepts in the system. Some theories posit that speech and music are understood largely due to their interrelated or systematic conceptual structure (Collins and Quillian, 1969; Lakoff, 1987; Potts et al., 1989). For interrelated concepts, other concepts within the system affect the intension (internally represented meaning relationships) of a given concept (Johnson-Laird, 1983). Moreover, the systematicity between related concepts allows for generalization using intensionality beyond similarity (e.g., Martin and Billman, 1994). Thus the difference in these classes of concepts depends on the systematicity of the interrelationships within the set of concepts. Isolated concepts lack this systematicity and therefore auditory objects that are not linked systematically should have little perceptual effect on each other.

For interrelated concepts, systematicity can be thought of as providing a virtual context that could influence the perceptual experience of auditory objects. Previous research has shown that although the general population does not possess the ability to label tones without the aid of a reference tone, they do demonstrate some sensitivity to the correct tuning of familiar music based on their long-term experience with music. For example, individuals tend to hum or sing songs at or near the original key in which they heard them (Levitin, 1994; Bergeson and Trehub, 2002). Individuals are also able to determine above chance if a familiar song is transposed one or two semitones (Terhardt and Ward, 1982; Terhardt and Seewann, 1983; Schellenberg and Trehub, 2003). Additionally, Smith and Schmuckler (2008) have demonstrated that non-musicians without absolute pitch performed better than chance at determining if an exceedingly familiar dial tone had been pitch shifted or not. These studies suggest that at least to some extent individuals store absolute pitch information as part of their detailed representations of specific auditory experiences, such as a frequently heard melody or dial tone. It is unclear, however, if this pitch information is abstracted from these specific experiences to form a categorical representation for generalized note knowledge in long-term memory (Posner and Keele, 1968; Goodman, 1972; Reed, 1972; Barsalou, 1983; Murphy and Medin, 1985). If this the case, then effects of such knowledge should be seen on stimuli that the listener has not heard before, such as isolated sinewave tones. More specifically, if the sensory trace for a given tone is disrupted due to backward masking, individuals would have to rely on category level knowledge in order correctly estimate the tone. As such, the error in people's estimates can be used to reveal the nature of underlying category information.

The current set of studies thus aims to explore the nature of isolated pitch perception and the degree to which it is guided by generalized note knowledge by using a tone adjustment task in which listeners hear one of several target tones backward masked by white noise followed by a starting tone. In the task, listeners were asked to adjust a starting tone's pitch to match the target tone's pitch. Because the target tone was backward masked by white noise, individuals had to rely on category knowledge in order to correctly estimate the tone given that the sensory memory for the target (or the echoic memory) was no longer available (Massaro, 1975). In order to determine if listeners' estimates are affected by generalized note knowledge, we asked listeners' to estimate tones from one of three different acoustic frequency distributions of target tones, which were all tones from the Western scale. If listeners do not possess generalized note knowledge to guide their estimations, their responses should be based solely on the local context and stimulus properties to the extent these are available after masking, for a given target such that their estimates are no different than randomly generated data (RGD). Randomly generated data can be produced by simulating responses drawn randomly from the set of possible frequency responses available on any given trial. The arbitrary responses are simply created by using a random number generator to select a value that corresponds to a tone within the stimulus distribution for any condition. This arbitrary response can then be subtracted from the true target tone location, similar to how response error is found for real participants. To adequately model random responses it is necessary to match the number of simulated subjects for each distribution to the number of participants for each distributional set. These randomly generated responses represent a model that assumes that a listener has no access to the representation of the actual target tone pitch given that it was masked, but represent the starting tone pitch and then generate random responses from that point irrespective of the target tone frequency. To the extent that the RGD models participant responses successfully, it suggests that listeners maintain no abstract representation of the target. To the extent that participant responses deviate from the model in terms of improved performance, this demonstrates the formation of an abstract representation of the target tone even after masking. Prior experiments suggest that individuals possess some degree of absolute pitch information for specific auditory experiences (Terhardt and Ward, 1982; Terhardt and Seewann, 1983; Levitin, 1994; Bergeson and Trehub, 2002; Schellenberg and Trehub, 2003). If this information is also abstracted from these experiences in the form of generalized note knowledge that can sufficiently impact isolated tone estimates, then we should find significant differences between listeners' tone estimates and the RGD.

The second experiment builds upon the results from the first study by examining the effect musical training and possession of true absolute pitch may have on the estimation of tones. There are extreme individual differences found in the population with regard to auditory expertise. In fact, the extreme individual differences in auditory expertise within the auditory domain, makes it particularly well suited to examine the impact of long-term prior knowledge in perception. For instance, a small portion of the population possesses absolute pitch (AP)—the ability to correctly identify an isolated musical note without the aid of a reference note. Any listener with absolute pitch presents an idealized case of a listener who should be unaffected by masking given that such listeners can immediately recode the fragile auditory target representation into a stable note category. Thus such listeners, by comparison to the RGD, present a standard of classification of isolated pitches based on the intensional structure of music rather than simple frequency-pitch auditory mapping. Additionally, individuals widely vary in the amount of musical training they receive. Musical training has been shown to be related to improvements in auditory and visual working memory (Chan et al., 1998; Brandler and Rammsayer, 2003; Ho et al., 2003; Jakobson et al., 2003, 2008;

Zafranas, 2004) as well as enhancements in attentional control (Hannon and Trainor, 2007). As such, the second experiment was designed to examine the differences in tone estimation for three different groups of listeners—absolute pitch (AP) possessors, non-AP individuals with matched musical expertise (ME), and non-AP musical novices (MN) on a micro-scale distribution where test tones differed by 20 cents and included two perceptually in tune notes (B4 and C5). Any differences in tone estimation found between AP listeners and ME should largely be due to explicit absolute pitch knowledge, while any differences between MN and ME should largely be due to music theoretic and music practice expertise.

EXPERIMENT 1

In order to examine the degree to which long-term knowledge influences the perception of isolated tones the present experiment used a tone adjustment task in which a target tone was presented at a specific frequency and then backward-masked with white noise. Backward-masking was used to reduce the availability of the sensory trace (or echoic memory) of the tone (Massaro, 1975) and instead rely on more abstract category level knowledge. Following the target tone and mask, listeners then heard a starting tone, which they were asked to adjust in frequency to match the pitch of the target tone (See **Figure 1**). Depending on the condition to which they were assigned, listeners were given target tones from one of three different distributions of the acoustic frequency of the tone stimuli (See **Figure 2**). The distributions of stimuli were constructed such that two distributions (Set 2 and 3) were a subset of the frequency range of the third distribution (Set 1). The manipulation of the frequency of the test sets was manipulated specifically to test for range effects in the tone matching judgments. For each trial the error or difference between listener's response and the actual test tone was measured. If listeners adjusted the starting tone, such that it was identical to the target tone, then was no error as the estimate was accurate. There were two kinds of errors that listeners could make; they could either over estimate or under estimate the target frequency. The error between the adjusted tone and the target tone was measured in 33-cent steps, as that was the smallest step size by which participants could traverse the distributions.

As previously mentioned, listeners were given target tones from one of three different distributions of stimuli. If listeners' specific auditory experiences are not abstracted in the form of generalized note knowledge, their responses should be based solely on the local acoustic frequency range context and the stimulus properties of the given stimulus tones. We modeled these effects using RGD, which showed that truly random frequency estimates would reflect the distribution of tested target tones. More specifically, random frequency estimates for lower pitched targets of a particular frequency distribution should on average show frequency over estimation, given that the probability of randomly selecting a tone higher than the target is greater than randomly selecting a probe response lower than the target due to frequency range limitations on the responses. Similarly, a random distribution of frequency estimates for higher pitched targets of a specific frequency distribution should on average show frequency underestimation, given that the probability of randomly selecting a response tone lower in frequency than the target is greater than randomly selecting a response probe tone frequency higher than the target. The most central members of a distribution should on average show zero error, as there would be an equal probability of randomly selecting a response probe tone that is either higher or lower in frequency than the target. By extension, stimulus sets with a frequency range that is larger should on average have greater overall error than stimulus sets with more restricted frequency ranges. The degree to which listeners' estimates reflect this response error pattern suggests the degree to which listeners' estimates are consistent with a random response model governed

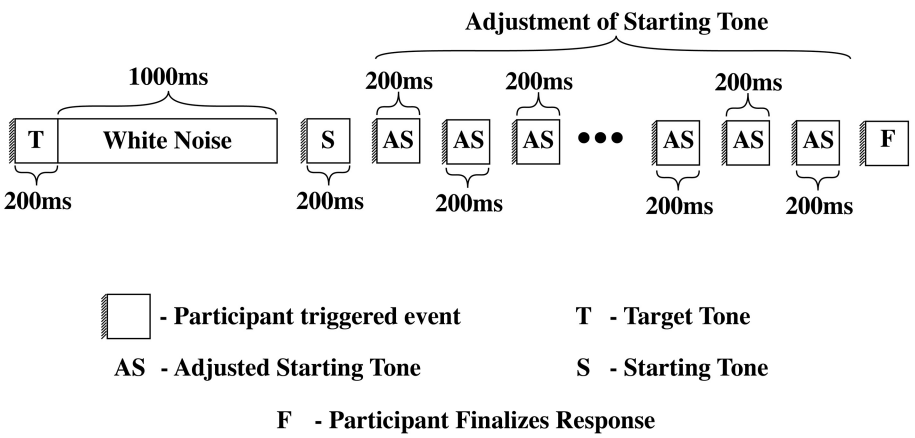


FIGURE 1 | On each trial, participants heard the target tone by clicking a circle with a T on it. This tone lasted 200 ms and was followed by 1000 ms of white noise. Individuals then heard the starting tone by clicking a circle with an S on it. There was no time limit between when the subjects heard the target tone to when they clicked the circle with an S on it to hear a starting tone. Individuals were then asked to adjust the starting tone to match the target tone. Participants adjusted the starting tone, traversing the stimulus series, either in 66-cent increments or 33-cent increments. Individuals were given as much time as they needed to adjust the starting tone. When participants were satisfied with their adjustment, they confirmed their answer by pressing a circle with a C on it.

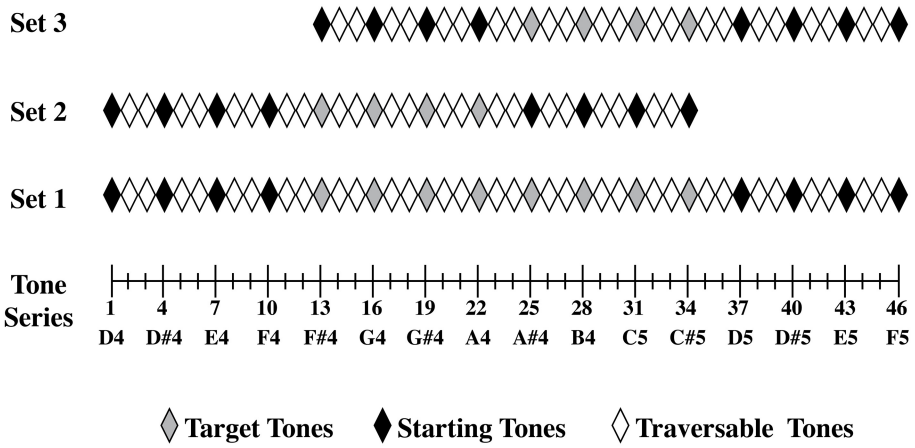


FIGURE 2 | Three sets of tones (or distributions of tones) were constructed from the original pure tone series that ranged from [D4] to [F5]. Set 1 consisted of stimuli 1–46 from our pure tone series, set 2 consisted of stimuli 1–34 from our pure tone series, and set 3 consisted of stimuli 13–46 from our pure tone series. In set 1, stimuli 1, 4, 7, 10, 37, 40, 43, and 46 were used as starting tones, while stimuli 13, 16, 19, 22, 25, 28, 31, and 34 were used as target tones. In set 2, stimuli 1, 4, 7, 10, 25, 28, 31, and 34 were used as starting tones, and stimuli 13, 16, 19, and 22 were used as target tones. In set 3, stimuli 13, 16, 19, 22, 37, 40, 43, and 46 were used as starting tones and stimuli 25, 28, 31, and 37 were used as target tones.

primarily by local stimulus context and target tone frequency. By contrast, if individuals possess musical note knowledge—in other words can identify the note category of a particular target tone–based on long term listening experience (Terhardt and Ward, 1982; Terhardt and Seewann, 1983; Levitin, 1994; Bergeson and Trehub, 2002; Schellenberg and Trehub, 2003) and this knowledge affects the frequency estimates of simple isolated target tones then we should find significant differences between listeners' tone estimates and the randomly generated responses.

METHODS

Subjects

Twenty nine undergraduates (18 male) were recruited from the University of Chicago undergraduate community and were between 18 and 26 years of age 1. While participants were not specifically recruited for their musical background, individuals reported to have studied or played an instrument (piano, violin, bass, guitar, flute, or singing) on average 5.5 years (*SD*: 4 years; Set1—M:3.6 years, *SD*: 3.3 years; Set2—M:6.3 years, *SD*: 4.7 years; Set3—M:6.5 years, *SD*: 4.2 years). Participants were either granted course credit or paid for their participation in the experiment. All participants had no reported history of either a speech or a hearing disorder. Additionally, informed consent, using a form approved by the University of Chicago Institutional Review Board, was obtained from all subjects.

Stimuli

A pure sinewave tone series ranging from [D4] to [F5] was generated using Matlab. (See **Figure 2** for the range of stimuli.) All stimuli were 200 ms in duration, were RMS normalized to 75 dB SPL, and had a sampling rate of 10 kHz with 16-bit samples. The lowest tone in the series, at the [D4] end of the series, had a frequency of 293.64 Hz. For each succeeding tone in the series, the frequency was increased by one third of a semitone or 33 cents. A step size of 33 cents was chosen as it is well above most listeners' thresholds for detecting pitch differences (e.g., Hyde and Peretz, 2004), while making the task challenging. The highest tone of the series, at the [F5] end of the series, the tone had a frequency of 698.39 Hz. The frequencies of the sine tones used were based on an equal tempered scale using tempered intervals. The masking noise was random Gaussian white noise and was generated in Matlab. Similar to the other stimuli, the white noise also was RMS normalized to 75 dB SPL and had a sampling rate of 10 kHz with 16-bit samples. The white noise sample however, was 1000 ms in duration.

Procedure

The experiment consisted of a tone adjustment task in which a target tone was backward-masked with white noise and then matched by varying the frequency of a starting tone. On each trial, participants were asked to click a circle with a T on it to hear a target tone, followed by one second of white noise. Individuals were then asked to click a circle with an S on it to hear a starting tone. There was no time limit between when the subjects heard the target tone to when they clicked the circle with an S on it to hear a starting tone allowing listeners to pace the experiment comfortably. Individuals were then asked to adjust the starting tone to match the target tone. Participants adjusted the starting tone, traversing the stimulus series, by clicking either big or small arrows located above and below the circle with an S on it. Arrows above the circle allowed participants to move higher in frequency in the series, increasing the starting tone's frequency, while the arrows below the circle allowed participants to move lower in frequency in the series, decreasing the starting tone's frequency. The larger arrows modified the starting tone in 66-cent increments, while the smaller arrows modified the starting tone in 33-cent increments. Participants were told to use the larger arrows to quickly move through the series and then to use the smaller arrows to make fine grain adjustments to their answer. Individuals were given as much time as they needed to adjust the starting

tone. When participants were satisfied with their adjustment and believed it matched the original target tone, they pressed a circle with a C on it to confirm their answer. Participants were then asked to press Space bar to continue to the next trial. All key press responses were recorded. The experiment was conducted binaurally over sennheiser HD570 headphones. **Figure 1** depicts the event structure for a given trial.

Participants were assigned to one of three stimulus distributions that varied the acoustic frequency range in the presentation of both the starting tones and target tones. Range variation was manipulated to determine the degree to which tone estimates were random as the current task was constructed so that the more variable individuals' estimates were, the more their estimates would reflect the distribution of tested target tones. Set 1 consisted of tones 1–46 from the pure tone series, set 2 consisted of stimuli 1–34 from the pure tone series, and set 3 consisted of stimuli 13–46 from our pure tone series. Set 2 and 3 are different subsets of Set 1 shifted in frequency range.

Figure 2 depicts how each distributional set was constructed and which tones were used as starting tones and target tones. All starting tones and target tones in each set were actual notes in the Western music 12-note chromatic scale. Each starting tone and target tone combination was presented two times each. Multiple starting tones were used across trials and counterbalanced in a pseudorandomized order to remove any general over- or underestimation of the tones due to the starting tone's position. Set 1 had 8 target tones and 8 starting tones, so there were 128 total trials. Set 2 and 3 each had 4 target tones and 8 starting tones, there were 64 total trials. 9 individuals were asked to adjust tones from set1, 9 individuals were asked to adjust tones from set 2, and 11 individuals were asked to adjust tones from set 3.

As RGD is distributionally specific, RGD was created for each distributional set. The arbitrary responses were created by using a random number generator to select a value that corresponded to a tone within the distribution (1–46 for the large distribution in Experiment 1, 1–34 for the smaller distributions in Experiment 1, and 1–27 for the micro distribution used in Experiment 2). We then subtracted this arbitrary response from the true target tone location, just as we did for the real participants. The number of simulated subjects for each distribution was matched to the number of participants for each distributional set. Therefore, 9 simulated random subjects were run for set 1, 9 simulated random subjects were run for set 2, and 11 simulated random subjects were run for set 3.

RESULTS

We calculated the frequency matching error for each adjustment trial by subtracting the actual target tone's stimulus number in the frequency series from the adjusted starting tone's stimulus number that was associated with the participant's confirmed response. The calculated adjustment error therefore represented the number of 33-cent steps by which an individual's adjustment was in error. Some of the adjustment errors were so extreme that it appeared that the participant entered an arbitrary response, by either failing to attend to the target tone on that given trial or by accidently confirming their adjusted response before they had made any adjustment. To find these outliers, we culled final responses that were greater or less than two standard deviations away from the average participant's response. This procedure was repeated for each test tone. Only 5.1% of responses for target tones from set 1, 6.4% of responses for target tones from set 2 and 5.1% of responses for target tones from set 3 were removed in this procedure.

The groups that corresponded to each of the three frequency ranges for targets did not significantly differ (alpha 0.05) in the amount of time that they took to adjust the starting tone [$F_{(2,\ 28)} = 0.303$, $p = 0.742$] to match the target tone. Participants who adjusted tones from Distribution 1 took an average of 6 s (*SD*: 2.5 s); participants who adjusted tones from Distribution 2 took an average of 5.6 s (*SD*: 0.7 s); participants who adjusted tones from Distribution 3 took an average of 6.4 s (*SD*: 3 s). Overall, individuals who adjusted tones from distribution 1 finished the task within 50–55 min, while individuals who adjusted tones from distribution 2 and 3 finished the task within 25 min, as distributions 2 and 3 had half as many trials.

In order to examine the impact of generalized note knowledge on the perceptual judgments of isolated tones the amount of matching error was found for each target tone, for each of the three sets of tones tested. If individuals used generalized note knowledge in the perception of tones then we would predict all target tones should have similar amount of error. This is not the case. Three separate One-Way repeated measure ANOVAs with Target Tone as the main factor was carried out for each distribution set (group of listeners) for the dependent measure of error. For each distribution, the effect of Target Tone was significant [Set 1, $F_{(7,\ 56)} = 21.255$, $p < 0.0001$; Set 2, $F_{(3,\ 24)} = 17.168$, $p < 0.0001$; Set 3, $F_{(3,\ 30)} = 23.722$, $p < 0.0001$] indicating that the amount of error for at least one test tone out of each series was significantly different.

Figures 3–5 plot the mean amount of error in 33-cent steps for each of the target tones for each of three distribution sets. Pairwise comparisons among the estimated marginal means were also performed using a Sidak adjustment. The significant (alpha 0.05) pairwise comparisons from these analyses are additionally shown in these figures. The RGD for each distribution set is additionally shown for comparison purposes in **Figure 6**. For the RGD data from set 1, all tones were significantly different from one another (using a Sidak adjustment—alpha 0.05) except tone 1 with 2, 3, and 4; tone 2 with 3 and 4; tone 3 with 4 and 5; tone 4 with 5, 7 and 8; tone 5 with 6 and 7; tone 6 with 7, and tone 7 with 8. In set 2 all tones in the RGD were significantly different from one another (using a Sidak adjustment—alpha 0.05) except tone 1 with 2; and tone 3 and 4. In set 3 all tones in the RGD were significantly different from one another (using a Sidak adjustment—alpha 0.05) except tone 1 with 2; and tone 3 with 4.

Visual inspection of the data shows that individuals' estimates were highly influenced by the distribution they received. For each distribution, higher pitched items showed underestimation as the probability of randomly selecting a tone lower than it was greater than randomly selecting a tone higher than it, while lower pitched items showed overestimation as the probability of randomly selecting a tone higher than it was greater than randomly selecting a tone lower than it, and central items showed near zero error as the probability of randomly selecting a tone both higher

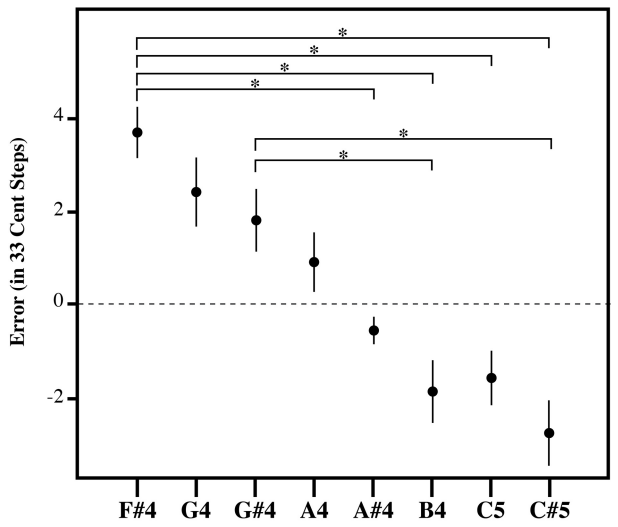


FIGURE 3 | The plot of the estimation error in 33-cent steps for each tone in Distribution Set 1. Significant pairwise comparisons (alpha 0.05) using a Sidak adjustment are indicated with an asterisk. Note that while individuals bisected the presented distribution at its midpoint, such that lower target tones in the series are overestimated and higher target tones in the series are underestimated similar to the randomly generated tone estimate data, individuals produced estimates that were significantly more accurate than the RGD (**Figure 6A**). Error bars represent ±1 s.e.m.

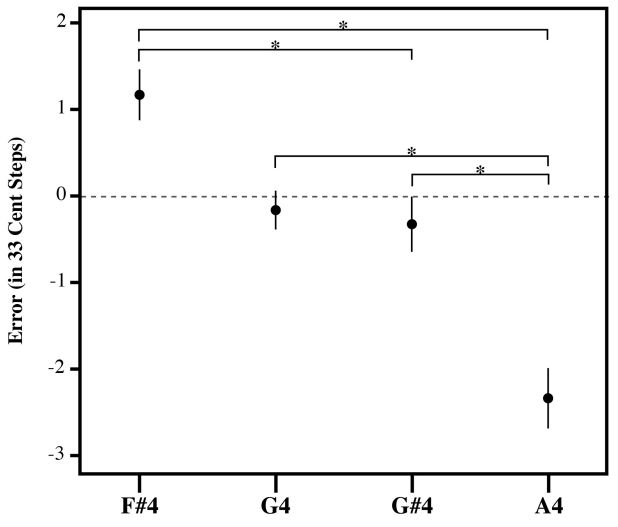


FIGURE 4 | The plot of the estimation error in 33-cent steps for each tone in Distribution Set 2. Significant pairwise comparisons (alpha 0.05) using a Sidak adjustment are indicated with an asterisk. Note that while individuals bisected the presented distribution at its midpoint, such that lower target tones in the series are overestimated and higher target tones in the series are underestimated similar to the randomly generated tone estimate data, individuals produced estimates that were significantly more accurate than the RGD (**Figure 6B**). Error bars represent ±1 s.e.m.

and lower than it was similar. In order to compare individual's estimates to the RGD, an error function was found for each participant (and each simulated subject in the RGD) by plotting the amount of error against the presented test tone series. Because the amount of error, when plotted against the target tone distribution used for each subject was linear in nature, a linear regression line was then fitted to each error function. From this, the x-intercept was calculated to infer the point of zero error. For distributional set 1 (the superset), the point of zero error was between stimulus 23 and 24, at 23.64. This point mirrors closely the actual midpoint of distributional set 1, which is between stimulus 23 and 24. For distributional set 2 (the lower frequency range), the point of zero error was between stimulus 16 and 17 at 16.86. Again, this is very similar to the actual midpoint of distributional set 2, which is between stimulus 17 and 18. For distributional set 3 (the higher frequency range), the point of zero error was at stimulus 30 at 30.00. This also closely echoed the true midpoint of distributional set 3, which is between stimulus 29 and 30. This indicates that individuals were highly sensitive to the presented distributions, suggesting that individuals' estimates were variable in nature.

While, individuals' general pattern of estimation error across the tones reflected a pattern consistent with context sensitive range-dependent estimates, listeners' responses may significantly and meaningfully differ in other ways from the RGD. For example, it is possible that while individuals' estimates were inexact, individuals did use their experience with Western music to inform their estimates. If this is the case, individuals should display significantly less absolute error across the distribution than found in the randomly generated estimates. To test this, the amount of estimation error was plotted against the presented test tone series for each subject. A linear regression line was fitted to each subject's estimation error function. This was also done for the RGD. The steepness of the fitted linear regression line was then used to assess the degree to which items were judged with more error, a steeper fitted regression line would necessarily denote more extreme overestimation of smaller items as well as more extreme underestimation of larger items in the series. As such, the slope corresponding to the fitted regression lines was used as a dependent measure in three separate planned independent sample T-tests (equal variances no assumed) to examine if individuals from each of the three distributions significantly differed from the RGD. Indeed, for each of the three distributions, individuals' responses showed significantly less error than the RGD [Set 1: $t_{(14.6)} = 7.456$, $p < 0.001$; Set 2: $t_{(11.7)} = 7.46$, $p < 0.001$; Set 3: $t_{(16.3)} = 3.88$, $p = 0.001$]. This suggests that individuals possess a limited amount of long-term pitch knowledge that helped to constrain their target tone pitch matching estimates.

DISCUSSION

The present results are similar to previous findings that suggest that individuals in the general population have some generalized absolute pitch knowledge (Terhardt and Ward, 1982; Terhardt and Seewann, 1983; Levitin, 1994; Bergeson and Trehub, 2002; Schellenberg and Trehub, 2003). Individuals' estimates of the frequency of isolated target notes are guided to some

extent by generalized note knowledge. For each tested distribution, individuals were significantly more accurate than the RGD. This suggests that listeners used abstracted note information that goes beyond the specific auditory context from which they were experienced, to form generalized note knowledge. However, while we found some evidence that long-term pitch knowledge helped to make estimates more accurate, individuals' estimates were still highly inaccurate. Pitch matching error for any given tone was largely dependent on the stimulus range in which it was presented. Individuals' pitch matching estimates were significantly influenced by the distribution of possible tones as they overestimated the lower pitched tones and underestimated the higher pitched tones in each distribution. Moreover, their estimates were most exact for the center of each tested distribution. Strikingly, this point of zero error directly mirrored the actual midpoint of each series. This sensitivity to the distribution is consistent with the idea that while individuals possess some generalized note knowledge, their estimates of isolated target notes are still variable in nature.

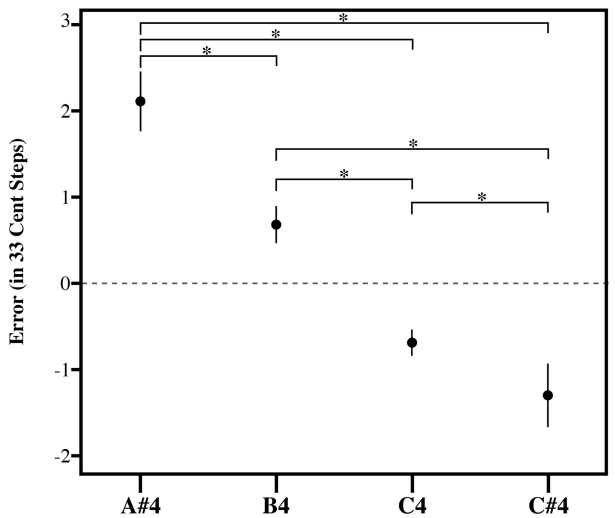


FIGURE 5 | The plot of the estimation error in 33-cent steps for each tone in Distribution Set 3. Significant pairwise comparisons (alpha 0.05) using a Sidak adjustment are indicated with an asterisk. Note that while individuals bisected the presented distribution at its midpoint, such that lower target tones in the series are overestimated and higher target tones in the series are underestimated similar to the randomly generated tone estimate data, individuals produced estimates that were significantly more accurate than the RGD (**Figure 6C**). Error bars represent ±1 s.e.m.

There are several possibilities as to why we failed to find strong evidence for latent note knowledge in the frequency estimates of target tones. One possibility is that prior pitch knowledge for individuals in the general population may be representationally sparse or underspecified due to poor encoding or insufficient experience. Another possibility is that individuals may not have recognized the notes in this experiment as examples of musical tones as they were pure sine wave tones. As such, they simply did not bring their latent note knowledge to bear on their estimates of these tones, since sine wave tones do not generally occur in every day musical experience.

While the general population does not have the ability to name a note without a reference note as guide, listeners with absolute pitch (AP) can do so and have well defined note categories. As such, this knowledge should affect their estimates of isolated tones. When people make use of category level knowledge, the most typical or central members of a category are best remembered, whereas items less typical or more extreme are distorted by this knowledge and are perceptually judged as being more typical or less extreme than they actually are. This effect is known as a central tendency effect. As such, Individuals with AP should make estimates that are influenced by their long-term absolute category knowledge. In this sense, the most typical members (perfectly tuned notes) should exhibit zero error, whereas mistuned notes within the category should be distorted by their category knowledge and be remembered as more typical then they actually are. The next experiment was conducted to further understand how such prior knowledge influences the estimation of isolated tones.

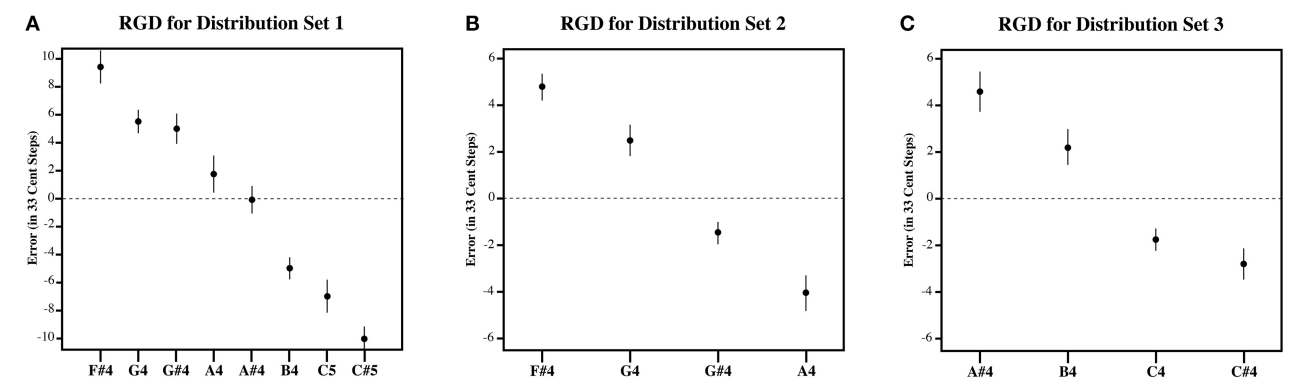


FIGURE 6 | The plot of the estimation error in 33-cent steps for each tone for the RGD. (A) Shows the RGD for set 1, **(B)** shows the RGD for set 2, and **(C)** shows the RGD for set 3. Note that the RGD bisects the presented distribution at its midpoint, such that lower target tones in the series are overestimated and higher target tones in the series are underestimated. Error bars represent ±1 s.e.m.

EXPERIMENT 2

In order to better understand how prior note knowledge affects the pitch estimates for isolated tones, Experiment 2 investigated whether differences in prior chromatic scale experience moderates the results reported in Experiment 1. In the present experiment, a fixed frequency range of stimuli was used with only two correctly tuned notes and 12 tones that were mistuned from those notes as target tones. The two correctly tuned notes were on either side of the center of the distribution so that prior pitch knowledge would be juxtaposed against a random or highly variable estimation pattern. To the extent that individual judgments are random, individuals' general pattern of estimation error across the tones should reflect a pattern consistent with context sensitive, range-dependent estimates. This is the case, as a model of randomly generated responses indicates that random estimates for lower frequency target tones of a distribution should on average be over estimated as the probability of randomly selecting a tone higher than it is greater than randomly selecting a tone lower than it, while random estimates for higher frequency items of a distribution should on average be underestimated as the probability of randomly selecting a tone lower than it is greater than randomly selecting a tone higher than it. Most importantly, even if individuals' estimates are random, there should be no estimation error for judgments at the center of the stimulus series despite the fact that the center is not actually a correctly tuned note. However, if prior note knowledge is abstracted to form generalized note knowledge that helps to inform individuals' pitch estimates for target tones, the two notes that are correctly tuned in the stimulus series should show reduced error. This is because the most typical or central members of a category should be best remembered, whereas items less typical or more extreme are distorted by this knowledge and are perceptually judged as being more typical or less extreme than they actually are. For this reason, listeners with absolute pitch (AP) should demonstrate less variability in their estimates. Thus, it is possible that individuals with more specific note knowledge (either explicitly in the for of AP knowledge, or more generally as more musical experience) will not show zero error for the mistuned center of the stimulus series. Instead, these individuals may show zero error for the targets that are the two correctly tuned notes. Further, neighboring tones that are slightly sharper than these in-tune tones should be underestimated, while tones that are slightly flatter than these in-tune tones should be overestimated.

However, it is also possible that AP listener judgments of notes will still show some pitch estimate error despite their note knowledge. A study by Hedger et al. (2013) provides clear evidence that AP perception is dependent on the tuning of recent experiences with particular notes and timbres, and that it is not reliant upon a direct or naïve realism framework (cf. Gibson, 1972), in which the underlying note is directly perceived. Given that the context of recent musical experience is important in the maintenance of note categories for AP listeners, AP listeners may show some error in their pitch estimates of isolated tones.

While it is clear that AP listeners have more extensive prior perceptual note knowledge than non-AP listeners, it is possible that AP listeners differ in other meaningful ways. For example, AP listeners will on average have more extensive music experience compared to the general population. There is a large body of literature that suggests that musical training is correlated with domain-general enhancements in cognitive processing. For example, music training is positively correlated with performance in auditory and visual working memory tasks (Chan et al., 1998; Brandler and Rammsayer, 2003; Jakobson et al., 2003, 2008; Ho et al., 2003; Zafranas, 2004) as well as with improved attentional control (Hannon and Trainor, 2007). Better auditory working memory or attentional control from musical training could help AP listeners perform better in the tone matching task by improved memory for target notes.

Therefore, we compared performance in a tone adjustment task for AP listeners with two additional groups–musical experts (ME) and true musical novices (MN). Schlemmer (2009) suggests that that ME may have richer generalized musical note knowledge than MN, as evidenced by a positive correlation between musical expertise and the ability to spontaneously sing a well-rehearsed piece on key without the aid of a reference tone. Overall though, note judgment differences between AP listeners and ME should be due to absolute pitch, while performance differences between MN and ME should be largely due to music theoretic and music practice expertise.

PARTICIPANTS

In order to understand how prior absolute pitch knowledge might influence the estimation of tones we recruited musical novices (MN), musical experts (ME) and Absolute Pitch listeners (AP) to take part in a tone-probe adjustment task similar to Experiment 1. Thirty-one individuals (11 musical experts, 4 females; 12 musical novices, 8 females; and 10 absolute pitch listeners, 6 females) participated in the experiment. The musical experts had studied or played an instrument (piano, violin, viola, cello, flute, singing) for at least 15 years (*M*: 23.1 years, *SD*: 7.7 years); all experts had training in the theory of harmony and in counterpoint during their studies. Musical novices had limited to no experience playing an instrument or singing (*M*: 2.9 years, *SD*: 3.2 years).

Absolute pitch listeners both identified themselves as possessing AP, but also passed a test that verified their ability to accurately produce isolated notes (for details, see the Procedure Section). Absolute pitch listeners, similar to musical experts reported substantial musical expertise, reporting to have studied or played an instrument (piano, violin, viola, cello, flute, singing) for at least 11 years (*M*: 22.1 years, *SD*: 9.9 years). All participants had no reported history of either a speech or a hearing disorder. Participants were either granted course credit or paid for their participation in the experiment.

Additionally, informed consent, using a form approved by the University of Chicago Institutional Review Board, was obtained from all subjects.

STIMULI

A 27-stimulus, pure sinewave tone test series ranging from a frequency that was 20-cent sharp [$B^{b}4$] to a 20-cent flat [$C^{\#}5$] was generated using Matlab. All stimuli were 200 ms in duration. For Stimulus 1, at the [Bb4] end of the series, the tone had a frequency of 471.58 Hz. For each succeeding stimulus in the series, the frequency was decreased by one tenth of a semitone or 10 cents. A

step size of 10 cents was chosen as it is toward the lower end of the range of most listeners' thresholds for detecting pitch differences (e.g., Hyde and Peretz, 2004), helping to make the task challenging even for AP possessors and ME. Consequently, for Stimulus 27, at the 20-cent flat [C#5] end of the series, the tone had a frequency of 547.99 Hz. The frequencies of the sine tones used were based on an equal tempered scale using tempered intervals.

The use of a finer grained distribution than in Experiment 1 allowed us to pit prior category knowledge against context-sensitive responses. If individuals' estimates are influenced by prior absolute pitch knowledge, then central tendencies or points of zero error should be observed for stimuli 9 and 19, as these tones of the series are, by Western music standards, perfectly tuned notes (B4 and C4). Tones near these perfectly tuned notes should be affected by prior absolute pitch knowledge, causing them to be remembered as more typical then they actually are. This means that slightly sharp notes will be underestimated, while slightly flat notes will be overestimated. Conversely, to the extent that individuals' estimates are variable, a point of zero error should be observed at or near stimulus 14, as this is the center of the tested distribution.

PROCEDURE

The experiment consisted of two parts. First, the participants were introduced to the stimuli of the experiment via a grouping task. The sole purpose of this task was to make certain that individuals understood the tones to be examples of musical notes, as the stimulus series only contained two perfectly in-tune notes. The grouping task therefore ensured that AP possessors, MN, and ME would use whatever prior note knowledge they have in perception of the tones. The grouping task was not necessary in Experiment 1 as all starting tones and target tones used in that experiment were actual notes in the Western music 12-note chromatic scale. In the grouping task all 27 stimuli appeared as clickable and moveable objects (gray squares) on a computer screen. Each object was marked by a random 3-digit number as an arbitrary label. For each participant, the stimuli appeared in a random order at the top of the computer screen to avoid any presentation ordering effects. Before beginning the task the subjects were told that they would have the opportunity to listen to and organize a set of sine wave tones. To indicate that these were indeed examples of musical tones even though a majority of the tones were not in tune notes, participants were also informed that the tones they would hear were examples of Bb, B, C and C# notes, but that some tones would be better examples of these notes than others. Subjects were asked to first listen to each stimulus by clicking on each object. They were then asked to sort the tones into the previously mentioned groups: Bb, B, C and C#. During this portion of the task the participant could hear each stimulus as many times as they wished in order to group the tones appropriately. Each subject took approximately 15 min to complete this portion of the experiment.

The second part of the experiment consisted of the tone adjustment task used in Experiment 1. The experiment was conducted binaurally over sennheiser HD570 headphones. Individuals adjusted tones from the set that they sorted in the grouping task. **Figure 7** shows how the set of tones was constructed and which tones were used as target tones and starting tones. Participants experienced each starting tone and test tone combination two times each, for a total of 80 trials. Multiple starting tones were used across trials in order to counterbalance, and thus remove any general over- or under-estimation of the tones due to the starting tone's position.

After the grouping and adjustment task, AP possessors completed a test of their AP ability. In a sound-attenuating booth, AP possessors were asked to sing or hum isolated notes, the names of which appeared one-at-a-time on a computer screen. Black key notes were produced eight times each—four times with the sharp symbol (#), and four times with the flat symbol (b)—while white key notes were represented four times each. There were thus 58 total trials. Participants could produce the notes in any octave they wished, and were instructed to hold a steady note for at least 2 s in order to accurately analyze the pitch.

In order to determine if individuals' estimates were better than RGD, simulations of the tone adjustment task were accomplished for the distributional set. As RGD is distributionally specific, RGD was created for each distributional set. The arbitrary responses were created by using a random number generator to select a value that corresponded to a tone within the distribution. We then subtracted this arbitrary response from the true target tone location, just as we did for the real participants. A total of 12 simulated subjects were run.

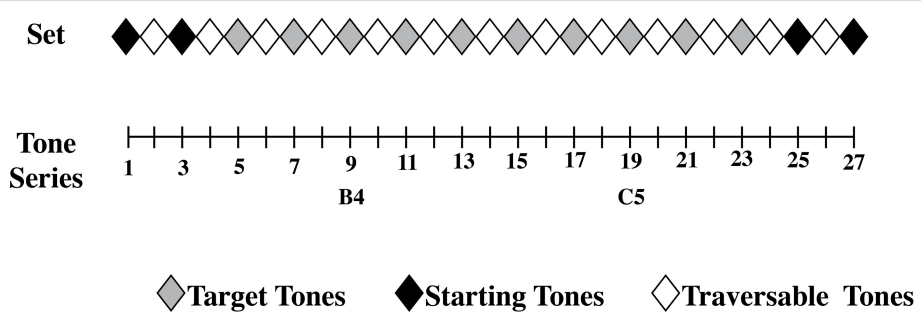


FIGURE 7 | The set of pure sinewave tones used for Experiment 2 ranged from a sharp [B4^{b}] to flat [C5$^{\#}$]. Stimuli 1, 3, 25, and 27 were used as starting tones, while stimuli 5, 7, 9, 11, 13, 15, 17, 19, 21, and 23 were used as test tones. Stimulus 9 was a perfectly in-tune [B4] and stimulus 19 was a perfectly in-tune [C5]. Stimulus 14, between stimuli 9 and 19, marks the central item of the tested distribution.

RESULTS

The same scoring and culling methods used in Experiment 1 was used in Experiment 2. Outlier responses were removed at a rate of 3.6% of responses for test tones for APP, 3.3% of responses for test tones for ME, and 2.1% of responses for MN were removed in the culling procedure.

The three groups did not significantly differ (alpha 0.05) in the amount of time that they took to adjust the starting tone to the final response on average across trials [$F_{(2,\ 32)} = 0.763$, $p = 0.475$]. AP possessors took an average of 7.2 s (*SD*: 4.9 s), ME took an average of 5.7 s (*SD*: 1.5 s), and MN took an average of 5.7 s (*SD*: 2.4 s). Overall, individuals were able to complete the task within 30–35 min.

As previously mentioned, the set of tones tested was specifically chosen to contrast prior category knowledge (two correctly tuned note targets) against the central tendency effect found in modeled RGD (the mistuned center of the stimulus series). If individuals' probe judgments are influenced by prior perceptual note knowledge, points of zero error should be observed for the two in-tune tones of the series (stimuli 9 and 19, B4 and C5. respectively). Additionally, neighboring out of tune stimulus tones that are slightly sharper than these in-tune tones should be underestimated while those tones that are slightly flatter than these in-tune tones should be overestimated. However, none of the groups' matching error responses reflected this pattern (see **Figures 8–10**). Instead, all three groups showed a point of zero matching error near the center of the tested distribution, which is not a correctly tuned note. This means that the lower pitched items of the test tone series showed positive error (or overestimation) and higher pitched items of the test tone series showed negative error (or underestimation). This suggests that all individuals' estimates were variable to some degree.

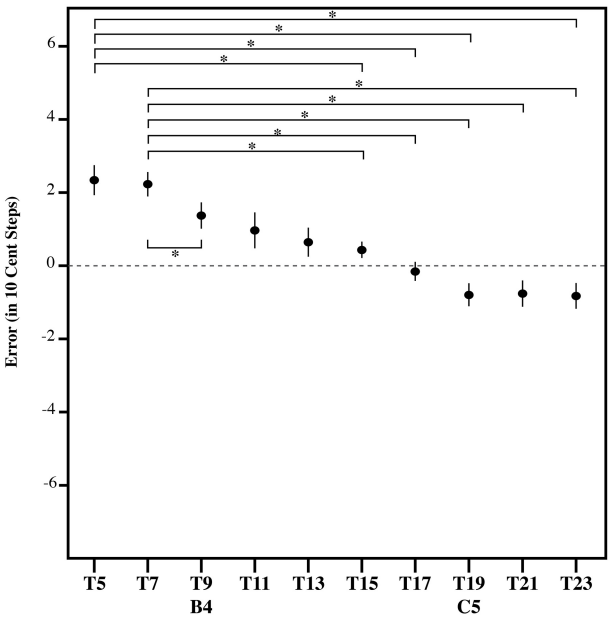


FIGURE 8 | The error in 10-cent steps for each tone in AP possessors. Significant pairwise comparisons (alpha 0.05) using a Sidak adjustment are indicated with an asterisk. Note that AP possessors bisect the presented distribution at its midpoint, such that lower tones in the series are overestimated and higher tones in the series are underestimated. However, note that this point for is closer to [C5] in AP possessors, than in ME (**Figure 9**), MN (**Figure 10**), and the RGD (**Figure 11**). Additionally, AP possessors (**Figure 8**) and ME (**Figure 9**), have flatter error patterns than MN and the RGD (**Figures 10**, **11**). Error bars represent ±1 s.e.m.

In order to determine if the amount of matching error for each test tone was different across the groups, we ran an omnibus repeated measures ANOVA with Target tone (10 different test tones were given as targets to match) as a repeated factor and Group (APP, ME, and MN) as a between subject factor. A significant main effect for Target tone was found [$F_{(9,\ 270)} = 38.01$, $p < 0.001$] indicating that the amount of error for at least one test tone out of the series was significantly different regardless of the listener group. Additionally, a significant main effect for Group was found [$F_{(2,\ 30)} = 5.208$, $p < 0.01$] denoting that musical experience or difference in prior pitch knowledge significantly altered the overall amount of matching error. Further, a significant interaction between Target tone and Group was found [$F_{(18,\ 270)} = 3.398$, $p < 0.001$] indicating that differences in musical experience or in prior note knowledge did not just globally lead to better or worse performance, but that they changed the judgments for each tone differentially.

To further examine the effect of target tone for each listener group, we carried out three separate simple effects One-Way ANOVAs, one for each group (AP, MN, and ME). Each of these One-Way ANOVAs had a significant main effect for Target tone. [For AP, $F_{(9,\ 81)} = 13.13$, $p < 0.001$; for ME, $F_{(9,\ 90)} = 11.91$, $p < 0.001$; for MN, $F_{(9,\ 99)} = 18.60$, $p < 0.001$]. This suggests that for all three groups, at least one target tone out of the series was significantly different. **Figures 8–10** plot the amount of error in 10-cent steps for each of the test tones for each group. Pairwise comparisons among the test tones were also performed using a Sidak adjustment. The significant (alpha 0.05) pairwise comparisons from these analyses are additionally shown. The RGD is additionally shown for comparison purposes in **Figure 11**. Upon visual inspection, all three groups show a pattern of error that is congruent with the idea that individuals' estimates were to some degree variable. Higher pitched items show underestimation, lower pitched items show overestimation and the central item (the mistuned center of the stimulus series) show near zero error.

However, a significant interaction in the omnibus Anova between Target tone and Group [$F_{(18,\ 270)} = 3.398$, $p < 0.001$] suggests that the pattern of error differed across the groups. One possibility is that while AP possessors' estimates were variable to some degree, their estimates may be still be influenced by note knowledge. For example, it is possible that the point of zero error in individuals with AP may be influenced toward one of the in-tune notes. This is because [B4] and [C5] may differ in familiarity, as C is a much more commonly experienced key signature than B (Simpson and Huron, 1994; Ben-Haim et al., 2014). If this is the case, than it is far more likely for us to see the point of zero error shifted toward C. To test for this, a linear regression line was fitted to each subject's error pattern (and each randomly generated

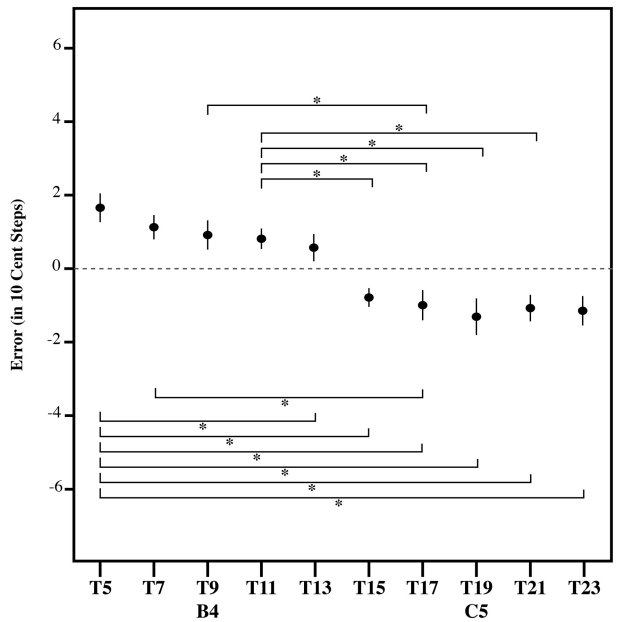


FIGURE 9 | The error in 10-cent steps for each tone in ME. Significant pairwise comparisons (alpha 0.05) using a Sidak adjustment are indicated with an asterisk. Note that ME bisect the presented distribution at its midpoint, such that lower tones in the series are overestimated and higher tones in the series are underestimated. However, note that ME (**Figure 9**), along with AP possessors (**Figure 8**) have flatter error patterns than MN (**Figure 10**) and the RGD (**Figure 11**). Error bars represent ±1 s.e.m.

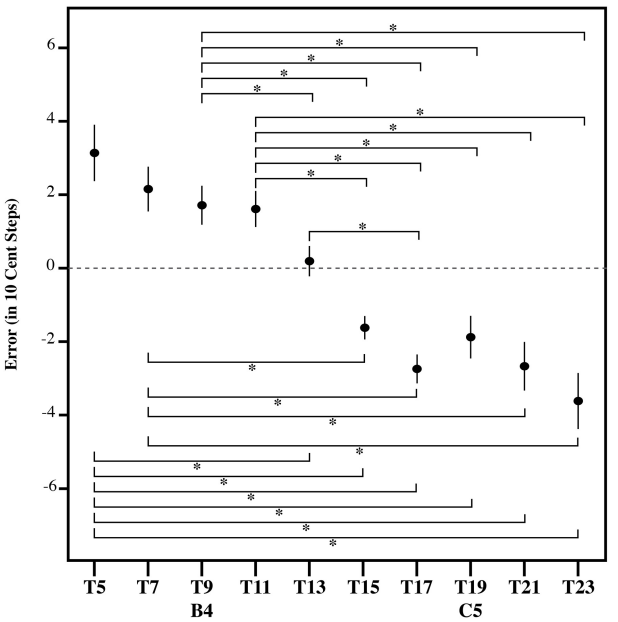


FIGURE 10 | The error in 10-cent steps for each tone in MN. Significant pairwise comparisons (alpha 0.05) using a Sidak adjustment are indicated with an asterisk. Note that MN bisect the presented distribution at its midpoint, such that lower tones in the series are overestimated and higher tones in the series are underestimated. However, note that MN (**Figure 10**) has a flatter error patterns than the RGD (**Figure 11**). Error bars represent ±1 s.e.m.

subject's error pattern), which was found by plotting the amount of matching error as a function of the test tone series. From this, the x-intercept was calculated to infer the point of zero error. To determine if any of the groups' x-intercept was significantly different than RGD, the RGD was added as a group. The location of the x-intercept within the test series was used as a dependent variable in a one-way analysis of variance examining the effects of Group (AP, MN, ME, and the RGD. A significant main effect of Group was found [$F_{(3,\ 44)} = 7.59$, $p < 0.001$] suggesting that at least one of the groups possess a significantly different central tendency location. Indeed, post hoc pairwise comparison testing using a Tukey HSD test showed that AP possessors' central tendency point was significantly different (alpha 0.05) than MN's, ME's, and the RGD's. More specifically, AP possessors' central tendency point was shifted away from the true center of the distribution, and toward [C5]. AP possessors' zero error point was near stimulus 16, while MN's, ME's, and the RGD's zero error point was near stimulus 14, the true center of the tested distribution (See **Figure 11**).

As previous mentioned, it is possible that differences in note knowledge or musical experience may affect the accuracy of pitch estimates for target tones. While individuals with more note knowledge may produce tone estimates that are inaccurate, they may be able to advantageously use long-term note categories to reduce such effects. If this is the case, AP possessors should display significantly less error across the distribution than MN and ME. However, it is also possible that domain general enhancements in working memory and attention, due to experience with Western music, may help individuals to better remember the isolated tones. If this is the case, then AP possessors and ME should display significantly less error across the distribution than MN. Additionally, it will be important to know how the error of individuals' tone estimates compared to the RGD. In Experiment 1, individuals from the general population, while variable in their estimates, showed significantly less error than the RGD. However, the current experiment used a distribution where only two of target tones differed by 20 cents and included two perfectly in-tune notes (B4 and C5). As such it is possible that some groups may not differ from the RGD in the amount of error they show across the distribution.

In Experiment 1, we argued that experience with Western note knowledge leads to less estimation error compared to random responses. If this is truly the case, we should find in Experiment 2 that differences in prior experience with the Western chromatic scale vary the amount of error in individuals' estimates of isolated tones. In order to determine if prior pitch knowledge affects pitch estimation for isolated tones, the amount of error was plotted against the presented test tone series for each subject. A linear regression line was fitted to each subject's estimation function (and to each simulated subject's estimation function). The steepness of the fitted linear regression line was then used to assess the degree to which items were influenced by the central

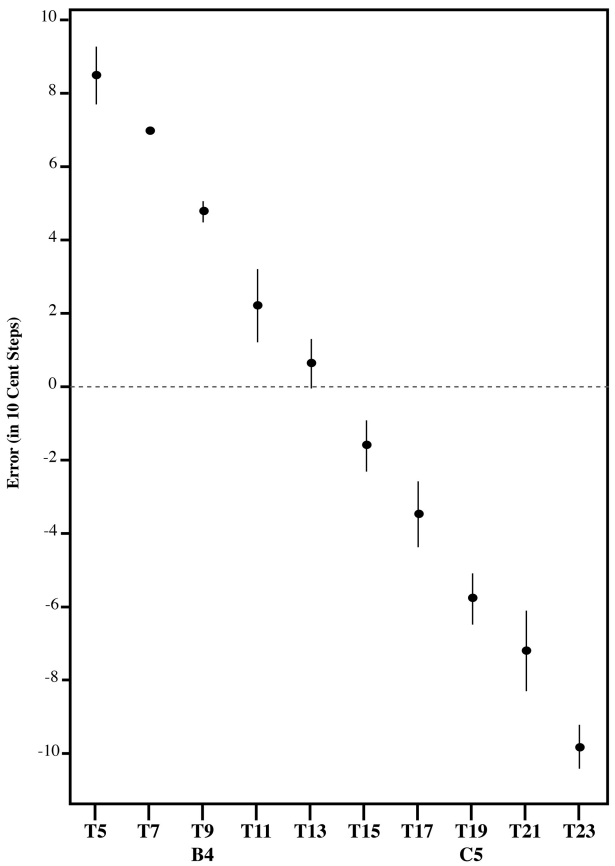


FIGURE 11 | The error in 10-cent steps for each tone in for the RGD. All tones are significantly different from one another using a Sidak adjustment (alpha 0.05) except tone 1 with 2; 3 with 4; 4 with 3, 5, 6, and 7; 5 with 6, and 7; 6 with 7 and 8; 7 with 8 and 9; 8 with 9 and 10; and 9 with 10. Note that the RGD bisects the presented distribution at its midpoint, such that lower tones in the series are overestimated and higher tones in the series are underestimated. However, note that the RGD (**Figure 11**) has a steeper error pattern than AP (**Figure 8**), MN (**Figure 10**), and ME (**Figure 9**). Error bars represent ±1 s.e.m.

items of the series as, a steeper fitted regression line would necessarily denote more extreme overestimation of smaller items as well as more extreme underestimation of larger items in the series. As such, the slope corresponding to each subject's fitted regression line was used as a dependent measure in a One-Way ANOVA with Group (AP, MN, ME, and the RGD) as the main factor. Indeed, a significant main effect of Group was found [$F_{(2,\ 32)} = 69.82$, $p < 0.001$]. *Post-hoc* pairwise comparison testing using a Tukey HSD test however, revealed that AP possessors' and ME's error patterns have a significantly flatter slope than the MN's and the RGD's error pattern. This suggests that domain general enhancements in working memory and attention, due to musical experience, helped AP possessors and ME better remember the isolated tones, and as such, helped to make estimates more accurate. Further, all individuals (AP, ME, and MN) showed significantly more accurate responses across the distribution than the RGD, indicating that all groups possess some long-term pitch knowledge that influences the estimation of isolated tones.

In order to verify that our AP possessors, who self-identified as having AP, did indeed possess the ability to produce an isolated note without the aid of a reference note, we analyzed the mean pitch of individuals' produced notes, comparing them to the objective standard used in Western music (A4 = 440 Hz). The pitch of each production was analyzed in Praat using the Burg algorithm (as reported by Press et al., 1992), and we used for analysis the latest possible window of 1 s where participants held a stable pitch. The reason we used the latest possible window for analysis is because only one participant had extensive vocal training, thus we wanted to allow individuals to adjust their initial vocal utterance to match their internal category standard if necessary. Overall, in addition to being self-identified as possessing AP, AP possessors were remarkably accurate at producing isolated musical notes, as the mean difference between their production and the objective tuning standard was less than half a semitone or 50 cents (*M*: −34.8 cents, *SD*: 30.5 cents, range: 21.9 to −45.5 cents). This is quite remarkable as the smallest distance between any two notes in the Western music scale is one semitone or 100 cents. Further, we never provided participants with feedback as to whether their sung note was correct (nor did we ever provide them with feedback throughout the entire experiment). Thus, all AP participants were well within an acceptable range for accurately producing isolated musical notes without the aid of a reference note. Individuals from the ME and MN did not participate in the pitch production task.

DISCUSSION

Despite a significant difference in musical experience and explicit note knowledge, the tone estimates of AP possessors, MN and ME had surprisingly similar patterns of error (see **Figures 8–10**). All three groups showed a point of zero error at or near the center of the tested distribution, suggesting that despite extreme differences in prior musical knowledge, individuals' estimates were still to some extent variable. Specifically, all three groups showed a zero error point associated with the center of the distribution such that lower pitched items of the test tone series were overestimated and higher pitched items of the test tone series were underestimated.

While all subjects were given exposure to the tones in the grouping task that preceded the tone matching task, it is still possible that the tone series was just too novel and as such did not bring prior note knowledge to bear on judgments of these tones. In this sense, it is possible that prior note knowledge might have had a more substantial effect on tone estimates if a more familiar timbre (e.g., piano) note series was used. Previous research indicates that AP possessors are more accurate and faster to identify notes when they are familiar with the timbre (Bahr et al., 2005; Schlemmer et al., 2005). Further Schlemmer (2009) has shown a positive correlation between experience with a particular piece and the ability to spontaneously sing it on key without the aid of a reference tone. Indeed, it is reasonable to assert that the use of prior pitch knowledge in the estimation of notes is likely modulated by the timbre, range and tonality of the notes used.

However, despite the probable novelty of the tone series' timbre, range and tonality, there were still notable differences in the pattern of error across the groups. Both AP possessors and ME had significantly less overall error than MN. This is consistent with the idea that domain general enhancements in working memory and attention, due to musical experience, helped individuals to better remember the isolated tones. It is also possible though that the differences between MN and those with musical expertise (which includes both AP possessors and ME) are not solely due to domain-general enhancements in cognitive processing. Previous work has shown that musicians possess a facility for the processing and memory of musical sounds, and that this facility is accompanied by enhancement and more diverse brain activity (Koelsch et al., 1999; Brattico et al., 2001; Gaab and Schlaug, 2003). Wickens (1973) has speculated that this enhanced and wider spread neural activation is reflective of a robust representational system that supports and improves the encoding of auditory events. As such it is possible that the differences between MN and those with musical expertise (which includes both AP possessors and ME) are not completely due to disparities in working memory capacity but also arise from differences in representational richness that exists between experts and non-experts.

Beyond demonstrating differences in the amount of variability in individuals' estimates, groups also differed in their point of zero error. More specifically, AP possessors' had a significantly different crossing point, than MN, ME, and the RGD such that AP possessors' point of zero error was near stimulus 16, closer to the in-tune [C5], while MN's, ME's, and the RGD's point of zero error was near stimulus 14, the true center of the tested distribution (See **Figure 12**). This is commensurate with the notion that prior note category knowledge such as found in AP possessors additionally influenced their estimates of tones. This was demonstrated with a shift in the distributionally based zero point error toward [C5], such that notes closer to [C5] had over all less error. Given that [C5] is a much more common note and key signature than [B4] (Simpson and Huron, 1994; Ben-Haim et al., 2014), the shift toward [C5] in AP subject's estimates provides additional evidence to a growing literature that AP possessors' note representations are based on the statistics of listening experience (Bahr et al., 2005; Schlemmer et al., 2005; Hedger et al., 2013).

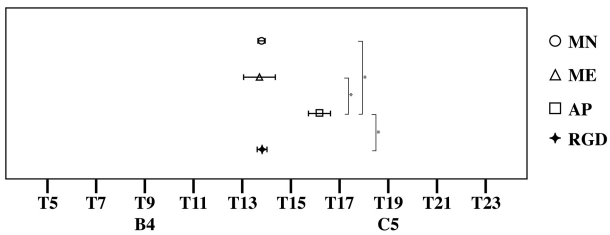


FIGURE 12 | The location of each group's point of zero error or central tendency point. Significant pairwise comparisons (alpha 0.05) using a Tukey HSD test are marked by an asterisk [*]. Note that AP possessors' central tendency point is shifted away from the true center of the distribution, and toward' [C5].

GENERAL DISCUSSION

Taken together, the results of these experiments provide evidence that musical novices (MN), musical experts (ME) as well as absolute pitch (AP) possessors all possess to some degree prior note knowledge as they all showed less error than the RGD, which is generated only on the basis of stimulus parameters. Further, the amount of prior note experience appears to modulate this error, such that more experience leads to more accurate tone estimates. In addition to these findings, possessing explicit note knowledge, as is the case for those with AP, appears to additionally influenced estimates.

Why would AP possessors, who have absolute pitch knowledge for note categories, demonstrate variability in their estimates that is not systematically related to their note categories? It is perhaps unsurprising that MN or ME make variable estimates, as they do not have explicit absolute pitch knowledge. However, for listeners with rich prior musical note category knowledge such as in AP, the memory of an isolated tone should be structured by long-term note categories. If this was the case, AP listeners should display significantly less error across the stimulus series than ME, who are matched on musical experience. This was not the case, as AP listeners, while showing less matching error than MN, showed similar amounts of matching error to ME. This suggests that the decrease in overall error is not due to robust AP note categories but to domain general enhancements in working memory and attention that stem from extensive musical experience. These results demonstrate that the category knowledge in AP is not as absolute as might be believed. Indeed, Hedger et al. (2013) demonstrated that changing the frequency tuning of notes in a musical piece quickly retunes the note category prototypes for AP listeners to be in accordance with the altered listening experience. Furthermore, this category shift generalized to notes not included in the detuned musical experience (albeit not to a different timbre), suggesting that Absolute Pitch perception is dependent on underlying statistical experience of tone frequencies. It is this generalization beyond the detuned notes experience that demonstrates strongly the systematicity of the note knowledge for AP listeners. The present data similarly suggests that Absolute pitch perception relies on an interaction between category knowledge and stability in listening experience.

From this perspective, the perception of auditory objects might be thought of as an active cognitive process given that perception occurs against the backdrop of prior experience. That is, even when simple tones are presented in isolation, individuals systematically perceive them in the context of prior musical note knowledge to a degree. The more experience one possess, the greater the influence of this knowledge on perception. In this sense, the perception of pitch, even in AP listeners should not be thought of as a simple template matching process. Clearly AP listeners do not directly access a note category from the frequency information in a tone. Rather, pitch information is perceived within the context of previous pitch experience. As such, the active use of prior pitch knowledge in the perception of simple, isolated tones prohibits a model of auditory perception that is

simply bottom up. Models of auditory perception should allow for the readjustment of subcortical processing, via the corticofugal system, to engage in egocentric selection, in which input from the brainstem is improved through feedback and lateral inhibition (Suga et al., 2002).

Overall, we have provided empirical evidence that all listeners possess to some degree prior pitch knowledge that affects the perception and subsequent judgments of isolated tones. Moreover, the amount of prior pitch knowledge modulated the degree to which estimates were accurate. Experienced listeners with substantial explicit knowledge and training showed less overall error than listeners without formal explicit training, suggesting that domain general enhancements in working memory and attention are associated with musical experience. Notably, all listeners showed less error than RGD. Additionally, listeners with absolute pitch showed a significant effect of note category knowledge over and above this musical experience. Lastly, the perceptual learning of the intensional structure of note categories does influence the estimates of isolated tones. These data suggest that auditory objects that have intrinsic relationships in pattern structure may be perceived under the influence of prior listening experience. This suggests that the extensional mapping of auditory objects as stimuli onto perceptual experiences follows a common set of principles in common with other psychophysical judgments. However, with sufficient perceptual training and experience, the systematicity of category knowledge can have an effect as well on the perceptual processing of these auditory objects, suggesting an active perceptual processing mechanism to instantiate such category knowledge.

AUTHOR CONTRIBUTIONS

Shannon L. M. Heald and Stephen C. Van Hedger designed the reported studies and collected data and analyzed the data together with Howard C. Nusbaum. Shannon L. M. Heald prepared the first draft and Stephen C. Van Hedger and Howard C. Nusbaum revised and both refined the manuscript to final form.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported in part by an ONR grant DoD/ONR N00014-12-1-0850, and in part by the Division of Social Sciences at the University of Chicago.

REFERENCES

Bahr, N., Christensen, C. A., and Bahr, M. (2005). Diversity of accuracy profiles for absolute pitch recognition. *Psychol. Music* 33, 58–93. doi: 10.1177/0305735605048014

Barsalou, L. (1993). "Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols," in *Theories of Memory*, eds A. C. Collins, S. E. Gathercole, M. A. Conway, and P. E. M. Morris (Hillsdale, NJ: Erlbaum), 29–10I.

Barsalou, L. W. (1983). *Ad hoc* categories. *Mem. Cogn.* 11, 211–227. doi: 10.3758/BF03196968

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behav. Brain Sci.* 22, 637–660. doi: 10.1017/S0140525X99532147

Ben-Haim, M. S., Eitan, Z., and Chajut, E. (2014). Pitch memory and exposure effects. *J. Exp. Psychol. Hum. Percept. Perform.* 40:24. doi: 10.1037/a0033583

Bergeson, T. R., and Trehub, S. E. (2002). Absolute pitch and tempo in mothers' songs to infants. *Psychol. Sci.* 13, 72–75. doi: 10.1111/1467-9280.00413

Brandler, S., and Rammsayer, T. H. (2003). Differences in mental abilities between musicians and non-musicians. *Psychol. Music* 31, 123–138. doi: 10.1177/0305735603031002290

Brattico, E., Näätänen, R., and Tervaniemi, M. (2001). Context effects on pitch perception in musicians and nonmusicians: evidence from event-related-potential recordings. *Music Percept.* 19, 199–222. doi: 10.1525/mp.2001.19.2.199

Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* 46, 369–384.

Chan, A. S., Ho, Y., and Cheung, M. (1998). Music training improves verbal memory. *Nature* 396:128. doi: 10.1038/24075

Collins, A. M., and Quillian, M. R. (1969). Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* 8, 240–247. doi: 10.1016/S0022-5371(69)80069-1

de Saussure, F. (1959/1916). *A Course in General Linguistics*. eds C. Bally and A. Schehaye, Trans ed Wade Baskin. New York, NY: McGraw-Hill.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316. doi: 10.1126/science.291.5502.312

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.* 23, 5235–5246.

Gaab, N., and Schlaug, G. (2003). Musicians differ from nonmusicians in brain activation despite performance matching. *Ann. N.Y. Acad. Sci.* 999, 385–388. doi: 10.1196/annals.1284.048

Genovesio, A., Brasted, P. J., Mitz, A. R., and Wise, S. P. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron* 47, 307–320. doi: 10.1016/j.neuron.2005.06.006

Gibson, J. J. (1972). "Outline of a theory of direct visual perception," in *The Psychology of Knowing*, eds J. R. Royce and W. W. Rozeboom (New York, NY: Gordon & Breach), 215–240.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Mem. Cogn.* 24, 608–628. doi: 10.3758/BF03201087

Goodman, N. (1972). "Seven strictures on similarity," in *Problems and Projects*, ed N. Goodman (New York, NY: Bobbs Merrill), 437–447.

Hannon, E. E., and Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends Cogn. Sci.* 11, 466–472. doi: 10.1016/j.tics.2007.08.008

Hedger, S. C., Heald, S. L., and Nusbaum, H. C. (2013). Absolute pitch may not be so absolute. *Psychol. Sci.* 24, 1496–1502. doi: 10.1177/0956797612473310

Ho, Y. C., Cheung, M. C., and Chan, A. S. (2003). Music training improves verbal but not visual memory: cross-sectional and longitudinal explorations in children. *Neuropsychology* 17:439. doi: 10.1037/0894-4105.17.3.439

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hyde, K. L., and Peretz, I. (2004). Brains that are out of tune but in time. *Psychol. Sci.* 15, 356–360. doi: 10.1111/j.0956-7976.2004.00683.x

Jakobson, L. S., Cuddy, L. L., and Kilgour, A. R. (2003). Time tagging: a key to musicians' superior memory. *Music Percept.* 20, 307–313. doi: 10.1525/mp.2003.20.3.307

Jakobson, L. S., Lewycky, S. T., Kilgour, A. R., and Stoesz, B. M. (2008). Memory for verbal and visual material in highly trained musicians. *Music Percept.* 26, 41–55. doi: 10.1525/mp.2008.26.1.41

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness (No. 6)*. Cambridge, MA: Harvard University Press.

Koelsch, S., Schröger, E., and Tervaniemi, M. (1999). Superior pre-attentive auditory processing in musicians. *Neuroreport* 10, 1309–1313. doi: 10.1097/00001756-199904260-00029

Lakoff, G. (1987). "Cognitive models and prototype theory," in *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, ed U. Neisser (Cambridge: Cambridge University Press), 63–100.

Levitin, D. J. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. *Percept. Psychophys.* 56, 414–423. doi: 10.3758/BF03206733

Mansouri, F. A., Matsumoto, K., and Tanaka, K. (2006). Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting Test analog. *J. Neurosci.* 26, 2745–2756. doi: 10.1523/JNEUROSCI.5238-05.2006

Martin, J. D., and Billman, D. O. (1994). Acquiring and combining overlapping concepts. *Mach. Learn.* 16, 121–155. doi: 10.1007/BF00993176

Massaro, D. W. (1975). Backward recognition masking. *J. Acoust. Soc. Am.* 58, 1059–1065. doi: 10.1121/1.380765

Murphy, G. L., and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol. Rev.* 92:289. doi: 10.1037/0033-295X.92.3.289

Nieder, A., Freedman, D. J., and Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science* 297, 1708–1711 doi: 10.1126/science.1072493

Posner, M. I., and Keele, S. W. (1968). On the genesis of abstract ideas. *J. Exp. Psychol.* 77:353. doi: 10.1037/h0025953

Potts, G. R., St. John, M. E., and Kirson, D. (1989). Incorporating new information into existing world knowledge. *Cognit. Psychol.* 21, 303–333. doi: 10.1016/0010-0285(89)90011-X

Press, W. H., Teukolsky, W. T., Vetterling, W. T., and Flannery, B. P. (1992) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cogn. Psychol.* 3, 382–407. doi: 10.1016/0010-0285(72)90014-X

Schellenberg, E. G., and Trehub, S. E. (2003). Good pitch memory is widespread. *Psychol. Sci.* 14, 262–266. doi: 10.1111/1467-9280.03432

Schlemmer, K. (2009). Das Gedächtnis für Tonarten bei Nichtabsoluthörern: Einflüsse von Hörhäufigkeit und musikalischer Ausbildung [Memory for tonality in non-absolute-pitch-possessors: influences of hearing frequency and musical education]. *Jahrbuch Musikpsychologie* 20, 123–140.

Schlemmer, K. B., Kulke, F., Kuchinke, L., and Van Der Meer, E. (2005). Absolute pitch and pupillary response: effects of timbre and key color. *Psychophysiology* 42, 465–472. doi: 10.1111/j.1469-8986.2005.00306.x

Simpson, J., and Huron, D. (1994). Absolute pitch as a learned phenomenon: evidence consistent with the Hick-Hyman Law. *Music Percept.* 12, 267–270. doi: 10.2307/40285656

Smith, N. A., and Schmuckler, M. A. (2008). Dial A440 for absolute pitch: absolute pitch memory by non-absolute pitch possessors. *J. Acoust. Soc. Am.* 123, EL77–EL84. doi: 10.1121/1.2896106

Suga, N., Xiao, Z., Ma, X., and Ji, W. (2002). Plasticity and corticofugal modulation for hearing in adult animals. *Neuron* 36, 9–18. doi: 10.1016/S0896-6273(02)00933-9

Terhardt, E., and Seewann, M. (1983). Aural key identification and its relationship to absolute pitch. *Music Percept.* 1, 63–83. doi: 10.2307/40285250

Terhardt, E., and Ward, W. D. (1982). Recognition of musical key: exploratory study. *J. Acoust. Soc. Am.* 72, 26–33. doi: 10.1121/1.387989

Wickens, D. D. (1973). Some characteristics of word encoding. *Mem. Cogn.* 1, 485–490. doi: 10.3758/BF03208913

Zafranas, N. (2004). Piano keyboard training and the spatial–temporal development of young children attending kindergarten classes in Greece. *Early Child Dev. Care* 174, 199–211. doi: 10.1080/0300443032000153534

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

Received: 01 April 2014; accepted: 03 August 2014; published online: 22 August 2014.
Citation: Heald SLM, Van Hedger SC and Nusbaum HC (2014) Auditory category knowledge in experts and novices. Front. Neurosci. ***8****:260. doi: 10.3389/fnins.2014.00260*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

REVIEW ARTICLE
published: 08 August 2014
doi: 10.3389/fnins.2014.00238

Emergence of category-level sensitivities in non-native speech sound learning

Emily B. Myers [1,2,3]*

[1] Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, CT, USA
[2] Department of Psychology, University of Connecticut, Storrs, CT, USA
[3] Haskins Laboratories, New Haven, CT, USA

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Matthew H. Davis, MRC Cognition and Brain Sciences Unit, UK
Caroline A. Niziolek, University of California, San Francisco, USA

***Correspondence:**
Emily B. Myers, University of Connecticut, 850 Bolton Rd., Unit 1085, Storrs, CT 06269, USA
e-mail: emily.myers@uconn.edu

Over the course of development, speech sounds that are contrastive in one's native language tend to become perceived categorically: that is, listeners are unaware of variation within phonetic categories while showing excellent sensitivity to speech sounds that span linguistically meaningful phonetic category boundaries. The end stage of this developmental process is that the perceptual systems that handle acoustic-phonetic information show special tuning to native language contrasts, and as such, category-level information appears to be present at even fairly low levels of the neural processing stream. Research on adults acquiring non-native speech categories offers an avenue for investigating the interplay of category-level information and perceptual sensitivities to these sounds as speech categories emerge. In particular, one can observe the neural changes that unfold as listeners learn not only to perceive acoustic distinctions that mark non-native speech sound contrasts, but also to map these distinctions onto category-level representations. An emergent literature on the neural basis of novel and non-native speech sound learning offers new insight into this question. In this review, I will examine this literature in order to answer two key questions. First, where in the neural pathway does sensitivity to category-level phonetic information first emerge over the trajectory of speech sound learning? Second, how do frontal and temporal brain areas work in concert over the course of non-native speech sound learning? Finally, in the context of this literature I will describe a model of speech sound learning in which rapidly-adapting access to categorical information in the frontal lobes modulates the sensitivity of stable, slowly-adapting responses in the temporal lobes.

Keywords: speech perception, phonetic category, second language acquisition, inferior frontal gyrus, superior temporal gyrus

INTRODUCTION

Phonetic categories, the basic perceptual units of language, are defined over distributions in acoustic space. For any phonetic category (e.g., /d/) there will be a range of acoustic tokens that will all be computed as acceptable members of a given phonetic category. To take a classic example, voiced and voiceless stops (e.g., /d/ vs. /t/) are primarily distinguished in initial position by the acoustic/articulatory parameter known as voice onset time, or VOT. For a native English speaker, VOTs less than about 30 ms are heard as /d/ sounds and those greater than 30 ms are perceived as /t/ sounds. The process of learning phonetic categories requires that the listener learn the boundaries of this acoustic space in order to understand how any given acoustic token maps to the phonology of his/her native language. To take the example given above, the English-learning child will learn that the voicing boundary falls at about 30 ms VOT in her language, but the Spanish-learning child will learn a boundary at about 0 ms VOT (Lisker and Abramson, 1964). This learning process is complicated by the fact that phonetic categories are typically defined by multiple acoustic parameters (e.g., VOT, vowel length, closure duration, burst amplitude). In this sense, we may think of the process of learning phonetic category boundaries as one of defining a hyperplane through multi-dimensional acoustic space.

In theory, all that is necessary for successful phonetic processing is the discovery of the location of phonetic boundaries in acoustic space. However, human speech perception is more complex than this. Over the course of development, acoustic differences that are contrastive in the child's native language become perceived as more distinctive, while those that are non-contrastive (i.e., they fall within the same phonetic category) become perceived as less distinctive (Eimas et al., 1971; Werker and Tees, 1999; Polka et al., 2001; Best and McRoberts, 2003; Kuhl et al., 2008). This perceptual pattern, namely excellent discrimination of items that fall between categories in the face of poor discrimination of items within phonetic categories, is referred to as categorical perception (Liberman et al., 1957). Through early childhood, this trajectory continues, with native-language contrasts becoming perceived more categorically and non-native contrasts becoming less categorical between ages 2 and 6 (e.g., Burnham et al., 1991). By the time listeners reach adulthood,

many phonetic categories are perceived categorically, and as such the mature phonetic processing system is not only sensitive to the boundaries of phonetic space, but exhibits perceptual warping such that certain portions of that space are easier to discriminate than others.

It is a matter of significant debate as to how categorical perception emerges. One proposal is that the statistical distribution of phonetic tokens in acoustic-phonetic space may provide sufficient information to reshape perceptual sensitivities even before functional phonetic categories have developed in the learner (Kuhl et al., 1992; Guenther and Gjaja, 1996; Maye et al., 2002, 2008). This view stems from the observation that the speech tokens that listeners are exposed to are not evenly distributed in acoustic space. For instance, the listener will hear many more examples of /t/ with a VOT near 60 ms will than with a VOT of 120 ms, although both are considered to be members of /t/ category (**Figure 1A**). Some evidence suggests that infant and adult listeners alike may be able to take advantage of distributional/statistical information in order to amplify acoustic distinctions that fall between different distributions and minimize those within the distribution (Maye et al., 2002, 2008; Hayes-Harb, 2007; Emberson et al., 2013). Crucially, this perceptual reshaping can happen even when listeners know nothing about the functional use of phonetic categories—that is, when listeners are only passively exposed to the input, and never hear speech sounds used referentially.

Nonetheless, young and old learners alike are exposed to additional sources of information regarding the sounds that are contrastive in their language. The use of phonetic categories to refer to different visual objects has been shown to result in better discrimination of those sounds (Yeung and Werker, 2009), and the appearance of different sounds in different lexical contexts may have a similar effect (Feldman et al., 2013). Ultimately, it is clear that the language learner must eventually learn the phonology of his or her own language. This sort of top-down information may continue to reshape perceptual sensitivities to these same sounds as the language user matures (**Figure 1A**). Given that the warping of perceptual space seen in adults may have arisen both as a consequence of passive, bottom-up data derived from the statistical distribution of tokens in the input, as well as the acquisition of functional, category-level information about the phonology of one's language, it is challenging to attribute behavioral and neural patterns we observe in adult phonemic perception to either bottom-up sensitivities to the acoustic input or top-down knowledge of phonetic category status.

This obstacle is particularly evident when discussing the neural systems that are responsive to phonetic category identity. For instance, if it is the case that statistical/distributional information in the signal is sufficient to guide the emergence of phonetic category identity, neural structures that are responsive to phonetic category structure may be those that have formed as a function of these bottom-up properties of the signal rather than as a response to the functional, linguistic use of phonetic categories. To the extent to which we believe passive mechanisms may also be sufficient to reshape sensitivities to complex acoustic information in auditory and auditory association cortex (Pallier et al., 1997; Zhou and Merzenich, 2007), mature, native language neural sensitivity may in large part reflect these bottom-up mechanisms.

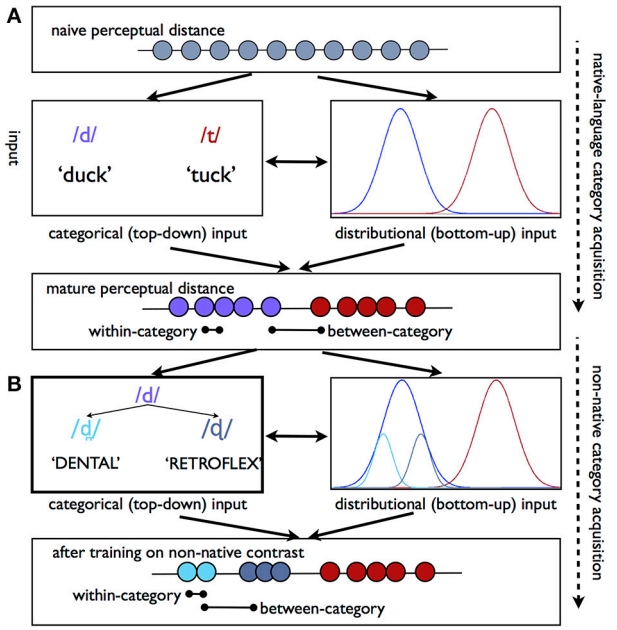


FIGURE 1 | Perceptual warping as a consequence of phonetic category learning. (A) Schematic of the process by which categorical perception emerges through development. Top line reflects the naïve perceptual distance between tokens along an arbitrary acoustic-phonetic continuum. Over the course of development categorical information (e.g., the use of tokens to refer to minimal pairs) and the statistical distribution of tokens in acoustic-phonetic space (e.g., more tokens are heard that fall near the center of the phonetic category) converge to warp perceptual sensitivities such that between-category contrasts are more perceptually distinct than within-category contrasts. **(B)** Non-native speech sound training paradigms primarily rely on categorical-level cues (e.g., explicit feedback), to reshape existing sensitivities. In this particular example, a listener must learn that two non-native sounds which are typically perceived as variants of /d/ correspond to different categories. This type of learning situation presents a particular challenge to the adult learner, given that the perceptual distance between these tokens in the mature listener is collapsed. Learning may proceed either via the top-down route, (left), or via passive exposure to statistical regularities in the input (right), or both. Over time, this information likewise results in differences for within- vs. between-category perceptibility.

In order to develop plausible hypotheses about the nature of phonetic category formation in adult, non-native acquisition, it is first important to discuss current evidence regarding the neural processing of native-language phonetic category structure.

NATIVE LANGUAGE PHONETIC CATEGORY STRUCTURE IN THE BRAIN

It has been well established that the bilateral superior temporal lobes are preferentially responsive to intelligible speech sounds compared to identifiable non-speech sounds (e.g., Belin et al., 2000, 2002), and compared to acoustically-matched sounds which are unintelligible as speech. (Okada et al., 2010; Evans et al., 2013). Recent evidence from direct cortical recording has revealed populations of neurons that code for dimensions of the phonetic

inventory, including place of articulation and manner of articulation, showing that the human temporal lobes are well-equipped to distinguish between the sounds of speech (Chang et al., 2010; Mesgarani et al., 2014). What is less clear is the extent to which these systems are specifically tuned to native-language contrasts or whether they show a more general sensitivity to, or preference for, many classes of speech sounds (for a more complete review, see Turkeltaub and Branch Coslett, 2010).

In order to answer this question, the review below is restricted to evidence in which the neural response reflects specific sensitivity to the internal structure of native-language speech categories. In particular, studies which show different responses to variability within and between categories can be said to show this kind of sensitivity.

MID-TO-POSTERIOR SUPERIOR TEMPORAL GYRUS TUNING TO NATIVE-LANGUAGE CATEGORY STRUCTURE

As a seat of complex acoustic processing, the bilateral temporal lobes play a primary role in processing the auditory details of the speech signal. Evidence suggests that there is a gradient of sensitivity along the temporal lobe from finer-grained acoustic processing near Heschl's gyrus (HG) to increasing specificity in tuning to one's native language as the processing stream flows in both the anterior and posterior directions along the STG/STS. In particular, middle portions lateral to HG have been shown to respond to native speech sounds compared to well-controlled non-speech sounds (Liebenthal et al., 2005; see Turkeltaub and Branch Coslett, 2010; DeWitt and Rauschecker, 2012 for meta-analyses). In contrast, regions including middle-STG territory lateral to HG and extending posterior along the STG/STS have been more tightly linked to phonological processing, and in particular have been shown to be sensitive to phonetic category structure. For instance, the bilateral superior temporal gyrus and superior temporal sulcus (STG and STS) are sensitive to how typical a speech sound is a member of its phonetic category (Guenther et al., 2004; Myers, 2007). This gradient response reflects the non-uniform structure of phonetic categories, suggesting that the temporal lobes are tuned to the internal perceptual structure of native-language categories, and are not merely sensitive to all speech sound dimensions.

The sensitivity of left posterior temporal areas in the perception of contrasts between- and within-category is supported by a series of studies using repetition suppression or habituation designs. While these studies differ in their details, all share a design in which a repeated presentation of a phonetic stimulus is followed by either an identical stimulus or a change in stimulus. Neural sensitivity to changes between and within the category are assessed by comparing activation for "change" trials to "repeat" trials. More categorical responses, as reflected by selective sensitivity to between-category compared to either repeated or within-category contrasts, were found in the left supramarginal gyrus, and in left posterior superior temporal sulcus (Joanisse et al., 2007; Myers et al., 2009).

Evidence that the temporal lobes respond to native-language contrasts also comes from the mismatch negativity paradigm. Larger MMN responses are seen to deviant tokens which cross a phonetic category boundary than those that change within the category (Phillips, 2001). Of interest, the MMN source is thought to arise from bilateral temporal cortex, shows greater left-lateralization for native language contrasts (see Naatanen et al., 2007 for review; Zevin et al., 2010), and MMN responses over the left temporal lobe are larger to phonetic than non-phonetic contrasts when employing direct cortical recording (Molholm et al., 2014), particularly in or near the STS. This MMN response is not restricted to temporal lobes however; the MMN response is thought to have a secondary source in left prefrontal cortex (Paavilainen et al., 2003, see further discussion of frontal contributions in section "Left inferior frontal involvement in categorical responses to native-language contrasts").

Discussion above has been limited to studies which specifically show differences in responsiveness to within vs. between-category contrasts. Nonetheless, converging evidence from other types of designs suggests that posterior portions of the left STG/STS are responsive to the category identity of native-language speech sounds (e.g., Desai et al., 2008; Chang et al., 2010; Liebenthal et al., 2010; Mesgarani et al., 2014). Of interest, speech category sensitivity in temporal regions is not limited to purely perceptual paradigms but it is also evident in auditory feedback for speech motor control. In particular, when speakers receive perturbations to auditory feedback that fall near the phonetic category boundary, greater compensation is seen in the speech production response, with concomitant greater activation for near-boundary compared to far-boundary shifts in the bilateral posterior STG (Niziolek and Guenther, 2013). Taken together, these results suggest that the posterior superior temporal lobes, particularly on the left, show fine-grained tuning to the acoustic properties of one's native language, with greater (or perhaps selective) neural sensitivity to acoustic distinctions that result in a change in phonetic category. It is of note that responses in the posterior STG/STS are not driven solely by bottom-up characteristics of the acoustic signal, but are also modulated by shifts in phonetic category boundary, and by changes in the perceptual status of the stimulus (e.g., non-speech to speech) (e.g., Desai et al., 2008; Gow et al., 2008; Myers and Blumstein, 2008).

LEFT INFERIOR FRONTAL INVOLVEMENT IN CATEGORICAL RESPONSES TO NATIVE-LANGUAGE CONTRASTS

While the temporal lobes no doubt shoulder much of the burden in processing the sounds of speech, evidence suggests that left prefrontal cortex also plays a role in the computation of phonetic identity. In two passive repetition suppression studies, responses to category-level information (e.g., greater responses to between-category than within-category shifts, yet no difference between within-category and repeated trials) were seen in premotor areas (Chevillet et al., 2013), and in an "invariant" response in the precentral gyrus and pars opercularis (Myers et al., 2009). Pre-motor areas which had been identified as sensitive to between-category changes showed significant task-related functional connectivity during passive listening to sites in the posterior temporal lobes (Chevillet et al., 2013), which led to the interpretation that phonetic category computations rely on forward projections between the temporal and frontal lobes along the dorsal route (Hickok and Poeppel, 2004). A recent analysis by Lee et al. (2012) examined category-level sensitivity of several brain regions using new

data in which participants passively listened to syllables along a ba—da continuum as well as using existing data from a repetition suppression paradigm (Raizada and Poldrack, 2007). In this study, the authors employed a moving searchlight technique with whole-brain multi-voxel pattern analysis (MVPA, Kriegeskorte et al., 2006) to search for clusters of voxels in which the patterns of activation could discriminate between two different phoneme categories (da vs. ba). Sensitivity to category-level information was seen in the left pars opercularis and pre-supplementary motor region as well as in the left superior temporal lobe. Converging evidence from studies in which cortical processing is disrupted using TMS also points to a role for frontal structures in computing category membership: stimulation of motor cortex sites slightly alters categorical perception in phoneme categorization and discrimination tasks (Mottonen and Watkins, 2009; D'Ausilio et al., 2012).

What is less clear is the precise role or roles of these frontal structures, which may indeed constitute functionally distinct subregions within the frontal lobes. The implication of premotor areas has led to the hypothesis that articulatory codes for speech may be activated to either guide perceptual hypotheses generated in the temporal lobes, or, more radically, to act as the contents of the abstract speech sound category (Liberman and Mattingly, 1985). At the same time, the influence of frontal areas may not be limited to access to articulatory information, nor, indeed, is category-sensitive activation limited to premotor cortex. Anterior to premotor cortex, regions in Broca's homolog have been found to be sensitive to category-level information in a domain-general sense, and evidence from single-cell recordings in non-human primates suggests that invariant responses to category membership may arise in frontal areas (e.g., Freedman et al., 2001). As such, the involvement of frontal areas may not reflect motor-related activity, but may reflect access to a more abstract category representation. In general, these results suggest that a complex of information arising from prefrontal regions generally may guide perception (Davis and Johnsrude, 2007; Liebenthal et al., 2013).

At the same time, the role of frontal structures in speech intelligibility "in the wild" has been questioned (Hickok and Poeppel, 2007; Hickok et al., 2011). It has been observed that lesions to left inferior frontal areas need not impair explicit decisions of phonetic category identity, and rarely create errors in phonemic perception (Basso et al., 1977; Rogalsky et al., 2011), and that while stimulation of premotor sites may impair categorization decisions, there is no evidence of deficits in comprehension as a result of such stimulation (Krieger-Redwood et al., 2013). Engagement of frontal structures for speech perception has been especially observed in the presence of ambiguity or noise in the signal (Binder et al., 2004; D'Ausilio et al., 2012), and as such frontal areas are argued to be peripheral to processing the sounds of speech. Some (D'Ausilio et al., 2012) while agreeing that frontal involvement for perception seems especially important in the context of noise in the signal, point out that noisy signals and imperfect productions are actually the norm rather than the exception in the typical language environment, and that we should resist the temptation to view frontal influences in speech perception as epiphenomenal. As such the types of activation patterns observed in studies of categorical perception can be accommodated by assuming that frontal structures are consulted in less optimal listening conditions.

Whether the codes accessed in the inferior frontal lobes are articulatory or abstract in nature, evidence suggests that coding in the left prefrontal areas is more categorical than that represented in the temporal lobe. This suggests an architecture whereby fine-grained acoustic-phonetic details of the speech stream are processed in the left STG/STS, and this information is then projected forward to prefrontal regions to consult with categorical-level codes in a complex of frontal areas (**Figure 2**).

NON-NATIVE PHONETIC CATEGORY ACQUISITION: A CASE OF FUNCTIONAL PLASTICITY

As discussed above, the mature language learner comes to the second-language learning process with a set of pre-established perceptual sensitivities which may either facilitate or hinder the acquisition of a new category. In particular, to learn a new phonetic contrast which falls within the acoustic territory occupied by native language sounds, the listener must learn to either (a) shift an existing phonetic category boundary, as in the case of an English speaker learning the a VOT boundary which corresponds to the Spanish d/t contrast, or (b) divide an existing phonetic category into two, as in the case of the English listener learning to perceive the Hindi dental vs. retroflex stop contrast (**Figure 1B**). This latter scenario seems particularly challenging, as an entire native-language architecture has developed which prevents the listener from perceiving distinctions within the phonetic category.

By the time adulthood is reached, one's sensitivities to native-language phonetic categories have reached a stability point. In fact, non-native categories, particularly those that fall within

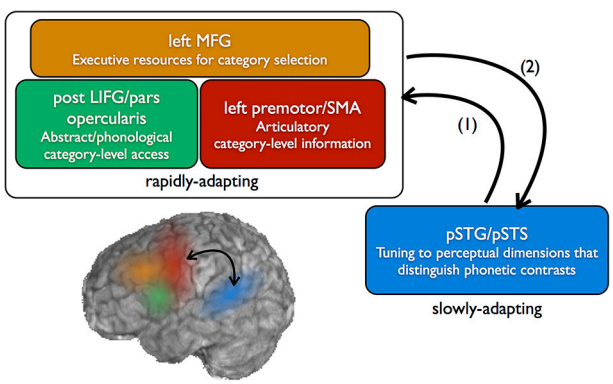


FIGURE 2 | Neural systems for the perception and learning of speech sound categories. Fine-grained sensitivity to acoustic dimensions that distinguish native speech sounds (e.g., VOT) is found in the posterior superior temporal gyrus (pSTG) and superior temporal sulcus (STS), which includes preferential sensitivity to speech categories, but, to a lesser degree, also sensitivity to within-category variation. In perception, sounds which are not well-categorized by this tuning (e.g., ambiguous sounds) feed forward to categorical-level coding in the frontal lobe (1). For non-native category learning which relies on top-down feedback, category sensitivities may emerge first in the frontal lobe, then feed back to posterior temporal areas to guide long-term changes in perceptual sensitivity (2). This frontal-to-temporal feedback corresponds to the top-down learning route shown in the bottom left portion of **Figure 1**.

an existing native-language category, are notoriously difficult to acquire in adulthood (Best et al., 1988). The fact that even motivated adults can struggle to distinguish certain non-native contrasts has led to conclusion that there is a critical period for phonetic category learning. This critical period may result from losses in neuroplasticity which prevent the adult listener from altering perceptual sensitivities in order to accommodate these into native language category structure (Pallier et al., 1997).

Nonetheless, with sufficient training, many individuals are able to learn to perceive non-native contrasts (Logan et al., 1991; Lively et al., 1993; Bradlow et al., 1997; Golestani and Zatorre, 2009), with some learners achieving native-like proficiency. Individuals who learn to speak a second language "in real life" (as opposed to in laboratory training conditions) have multiple sources of information which can guide the formation of new phonetic categories. Much as infants may be able to use information regarding the statistical distribution of phonetic tokens in acoustic space to reshape sensitivities, adults who are exposed to a non-native language will likely hear the same kinds of distributional information, whether they are able to take advantage of it or not (**Figure 1B**). Crucially for the adult learner, top-down information about phonetic category identity, either in the form of referential information (e.g., using two sounds to refer to two different words) or even through explicit classroom instruction, is often very salient in the environment. Unfortunately, almost all studies regarding the emergence of non-native phonetic sensitivity in the brain have used training paradigms where top-down information about category identity is provided to participants (see left side of **Figure 1B**). As such, we can draw limited conclusions regarding the emergence of neural sensitivities to non-native contrasts via more passive, bottom-up mechanisms in which listeners capitalize on distributional properties of the input.

PERCEPTUAL WARPING FROM NON-NATIVE CATEGORY TRAINING

Before considering the neural structures that are sensitive to phonetic category training, it is first important to assess whether top-down (e.g., categorization) training results in a perceptual pattern that resembles native language perception. As discussed above, acquisition of a native-language contrast appears to involve not only learning the boundary between categories, but also results in changes in perception of acoustic contrasts within and between these categories. Given that the types of training paradigms used in many studies bear a scant relationship to the authentic language acquisition environment, it would not be surprising to find that participants might successfully be able to complete a categorization task using non-native stimuli (that is, learn the location of the category boundary) while showing no difference in the relative perceptibility of between and within-category contrasts. Fortunately, converging evidence suggests that training participants on category-level information results in changes in discriminability of tokens across the trained continuum. Studies investigating training on the /l/ vs. /r/ contrast in native-Japanese listeners (McCandliss et al., 2002), and the Hindi dental vs. retroflex stop contrast (/d̪/ vs./ɖ/) in English listeners (Golestani and Zatorre, 2009) show that training on categorization tasks transfers to discrimination tasks, and specificity of the discrimination peak appears to be closely linked to both the location of the learned category boundary for each participant as well as to the relative success of each listener in acquiring the new contrast (Guenther et al., 1999; Wade and Holt, 2005; Golestani and Zatorre, 2009; Swan and Myers, 2013).

FUNCTIONAL BRAIN CHANGES RESULTING FROM NON-NATIVE CATEGORY TRAINING

When adults learn a non-native contrast, either via explicit category training, or from more naturalistic experience, brain structures which show specific sensitivity to native language contrasts must somehow reshape responses in order to accommodate a new categorical division of acoustic space. In general, we may ask whether the same neural resources are recruited for non-native speech sound perception following training as are implicated for native-language perception. Non-native phonetic training often takes the form of categorization training on either syllables or minimal pairs with explicit feedback to participants, often using a perceptual fading design, in which participants initially categorize maximally distinct tokens, then proceed to finer distinctions in a stepwise fashion (Golestani and Zatorre, 2004; Liebenthal et al., 2010; Myers and Swan, 2012). In this situation the availability of category-level information can be said to be at its maximum, as participants receive feedback regarding the accuracy of the categorical decision. When examining task-related activation before and after training, a wide network of regions are recruited, including bilateral temporal and left inferior frontal structures (Callan et al., 2003; Golestani and Zatorre, 2004) which show greater task-related activation to non-native sounds after compared to before training. Concordant evidence using a similar training paradigm yielded greater activation for non-native categorization post-training in a series of frontal regions and left inferior parietal regions (Ventura-Campos et al., 2013). Given the explicit nature of the categorization task, these studies are vulnerable to the criticism that the activation in inferior frontal regions is related to the metalinguistic task, rather than to the perception of phonetic category differences *per se* (see Section, "Left inferior frontal involvement in categorical responses to native-language contrasts," above).

Nonetheless, a study from our lab supported the involvement of a separate set of frontal structures, namely the left and right middle frontal gyri in categorical perception of learned speech sounds (Myers and Swan, 2012). In this study, participants were trained to categorize a three-way phonetic continuum (voiced stops ranging from dental to retroflex to velar place of articulation: /d̪/vs./ɖ/vs./g/) according to two different boundary locations, with one group trained to place the category boundary between the dental and retroflex tokens, and a separate group trained to place the category boundary between the retroflex and velar tokens. Participants were trained over two sessions, and neural sensitivity post-training was assessed using an short-interval habituation design which did not require participants to categorize speech sounds (see Joanisse et al., 2007; Myers et al., 2009). Despite the fact that the task required no judgments of phonetic category identity during scanning, activity in the bilateral middle frontal gyri reflected differential sensitivity to between vs. within-category contrasts according to the training of the participants. Of interest, no difference in activation for between

vs. within-category contrasts was seen in the temporal lobes, suggesting that differential responsiveness to learned category structure need not rely on retuning of sensitivities in the temporal lobe.

Support for the involvement of inferior frontal regions for non-native category learning can be seen in other passive paradigms. An analysis of resting-state functional data before and after intensive (one day) and distributed (six sessions) of non-native category training suggested that a decrease in degree of functional connectivity between two regions of interest in the left frontal operculum and left superior parietal lobule was significantly correlated with participant accuracy (Ventura-Campos et al., 2013). To unpack this result further, this suggests that individuals who were more successful in learning the non-native contrast showed a decrease in the degree of coherence between frontal and parietal structures, perhaps reflecting a decreased reliance on the frontal-to-parietal connection over the course of learning.

Nonetheless, training-related activity is not exclusive to these frontal regions. A series of training studies have shown significant involvement of temporal structures in sensitivity to trained speech and complex non-speech sounds. Liebenthal et al. (2010) trained participants over four sessions to identify non-speech sounds which resembled speech sounds in their spectral and temporal properties. Activation in the left posterior STS increased for trained non-speech sounds following training, with additional small clusters in left inferior frontal areas. Similarly, Leech et al. (2009) used an implicit training method which paired complex non-speech sounds with unique characters in a video game. After several sessions playing the game, the degree of increased activation within a speech-selective ROI in the left STS posterior to HG correlated with the degree of training success. Notably, this pattern did not emerge in a whole-brain analysis, and it may be the case that the creation of the speech-selective ROI may have eliminated the consideration of regions that would not respond to the speech vs. environmental sound contrast. Left posterior STS/STG activation has also been shown to correlate with training success in pitch pattern learning (Wong et al., 2007).

It is possible that the asymmetry between studies which have shown involvement of temporal regions in novel contrast sensitivity and those which have not may be attributed to the duration and/or intensity of training. Our study (Myers and Swan, 2012) employed only two 45-min sessions of training, whereas other studies have employed multiple intense training sessions. One proposal is that sensitivity to category-level information emerges early in the frontal lobe and only later is evident in temporal structures. This pattern would be consistent with a variety of proposals outside the language literature which suggest a shift from executive or category-level processing to sensory-based processing as expertise is gained (Ahissar and Hochstein, 2004; Nahum et al., 2008).

In order to address this question, we performed a replication of Myers and Swan (2012) in which we extended the training to ten 45-min sessions over 2 weeks (Myers et al., under review). Participants in this study were now trained to just distinguish dental and retroflex voiced stop consonants. Pre- and post-training scans were performed using the short-interval habituation design (Myers and Swan, 2012), and during scanning participants were asked to perform a pitch detection task in which they responded to high-pitched syllables on infrequent catch trials. Rather than search for areas which show global changes in activation as a function of training, we targeted regions which showed a differential sensitivity to between-category compared to within-category contrasts. Similar to other studies investigating categorical perception, the logic was that regions which showed sensitivity to the learned category structure following training could not be said to be influenced merely by changes in attention, motivation, or familiarity with the stimuli. At pre-test, only the left middle frontal gyrus showed differences in activation for between- compared to within-category stimuli. After training, activation differences were seen in a bilateral network including the left precentral gyrus, right and left STG, left IPL, and right insula. Importantly, both left and right posterior STG were shown to be correlated at post-test with participants' behavioral accuracy at post-test, suggesting that temporal activation resulting from 10 days of training was not only sensitive to the "categorical" nature of the stimuli (between vs. within) but also was predictive of learning.

INDIVIDUAL VARIABILITY IN SPEECH SOUND LEARNING

Many of the above-mentioned studies have searched for the neural correlates of variability in the perception of non-native contrasts. Variability in non-native perception is evident not only in training studies, but also in the varying degrees of proficiency that second-language learners attain (e.g., Bradlow et al., 1997; Flege et al., 1999). Studies which have examined the neural correlates of these differences among learners have come to differing conclusions regarding the source of this variability. Diaz et al. (2008) report that poorer perceivers of non-native contrasts showed an attenuated MMN response compared to better perceivers. The source of the MMN was inferred from the latency and distribution across electrodes, and was hypothesized by the authors to be the frontal component. The authors interpreted this response as reflecting engagement of an attentional network in better perceivers, whereas the lack of difference in the temporal component reflected similar fidelity in acoustic-phonetic processing across better and poorer percievers. By contrast, a study by Raizada showed that the patterns of activation within the right Heschl's gyri of Japanese L2 learners were predictive of that population's ability to discriminate /l/ vs. /r/ contrasts (Raizada et al., 2010). In the end, it is likely that functional variation at multiple points in the phonetic processing stream contribute to differences in learning success, with some learners excelling because of superior acoustic processing, and others achieving success due to the appropriate deployment of auditory attention, for instance.

Individual differences in brain structure are also predictive of phonetic learning success. Work by Golestani et al. (2002) and Golestani and Pallier (2007) showed that better learners showed differences in brain morphology in the left HG and a greater leftwards asymmetry in parietal cortex which was evident in WM volume. This asymmetry may reflect more efficient or precise coding of acoustic information which is especially relevant in speech sound learning (although see Burgaleta et al., 2014 for

a null finding relating brain morphology to speech perception abilities in a bilingual population). An advantage for processing the fine-grained aspects of sound might have surprising professional consequences as well. A unique study (Golestani et al., 2011) found that individuals who were employed as phoneticians showed differences in the morphology of left Heschl's gyrus compared to a control group. Of interest, there was also a correlation between the surface area and structure of the left pars opercularis and years of experience working as a phonetician, providing a hint that frontal differences in morphology may have arisen through experience-induced plasticity rather than from innate differences in brain structure.

This finding raises the question of whether experience learning a non-native phonetic contrast might actually induce structural changes in the brain. This type of plasticity is not unprecedented. Changes in brain morphology have been found following training on a variety of tasks (see Zatorre et al., 2012 for a review) and relevant for the current discussion, following a semester of intensive second-language learning (Stein et al., 2012). In our study of intensive non-native speech sound training (Myers et al., under review), changes in gray matter volume were seen in a region deep to the left supramarginal gyrus comparing pre-training scans to post-training scans. This same region is among the set of regions in which individual variation is associated with successful phonetic category learning (Golestani et al., 2007), and with individual differences in non-native sound production (Golestani and Pallier, 2007). Moreover, in our study, the coherence of white matter pathways (as measured by DTI) near the arcuate fasciculus in this same vicinity was seen to correlate with learning success, suggesting that the strength of frontal-to-posterior connections along the dorsal route contributes to non-native category learning. Taken together, these results suggest that even relatively short-term training can serve to strengthen connections that are necessary for non-native speech sound learning.

A FRONTAL TO TEMPORAL ROUTE FOR PHONETIC CATEGORY LEARNING

The extant literature on non-native speech sound learning suggests that the long-term consequence of speech category training is the retuning of posterior temporal regions such that they show increased sensitivity to the dimensions of the learned speech sounds. Of note, this same region also shows sensitivity to phonetic category structure in native speech perception which is presumably acquired slowly over the course of development. Broadly speaking, this is consistent with most models of the neural bases of speech perception (Hickok and Poeppel, 2007; Rauschecker and Scott, 2009). However, data suggests that short-term adjustments to learned phonetic category structure may be seen first in the frontal lobe (Myers and Swan, 2012), and only after sustained or more intensive training do these same sensitivities appear in the posterior temporal lobe (e.g., Leech et al., 2009; Myers et al., under review). Moreover, individual training success correlates with the coherence of white matter pathways at pre-training (Myers et al., under review), in an area that is consistent with the dorsal stream route connecting posterior temporoparietal regions to frontal structures (Hickok and Poeppel, 2007). Of note, this frontal-to-temporoparietal route is not the only connection which has been shown to correlate with non-native training success. Resting-state functional connectivity before and after training reflects a decreased reliance on frontal-to-superior parietal connections after training (Ventura-Campos et al., 2013) which has been attributed to a decreased reliance on a "salience" network. Of note, Ventura-Campos and colleagues also show strong resting-state connectivity between the frontal operculum and the SMG, but this connectivity did not show any significant correlation with training success. The authors speculate that this lack of correlation may in part reflect the lower individual variability shown in the frontal-to-SMG connectivity findings.

This pattern of results leads us to propose that early learning of non-native speech categories in the context of explicit top-down information involves first feed-forward connections from posterior temporal cortex to ventrolateral prefrontal cortex (Garell et al., 2013), where acoustic representations access category-level (articulatory, phonological, or abstract) information (**Figure 2**). Categorical sensitivity to non-native speech sounds emerges first in the inferior frontal lobe as participants learn the boundaries through acoustic space which define functional categories. This allows for rapid learning of category boundaries without fundamentally reshaping neural sensitivity to low-level details of the signal. Over time, frontal-to-temporal feedback connections may serve as an error signal on auditory sensitivities to these speech sounds, reshaping the sensitivity of auditory association cortex. The view that frontal-to-temporal feedback signals may play a role in rapid auditory plasticity finds support from animal models (Winkowski et al., 2013), and human data suggests that stimulation of frontal sites may facilitate auditory perceptual learning (Sehm et al., 2013). We suggest that the process of retuning sensitivities in the temporal lobes unfolds more slowly, over the course of minimally several days of training or experience.

Notably, our findings suggest that learners can achieve at least moderate success in training without any detectable change in the responsiveness of the temporal lobes (Myers and Swan, 2012). One open question is whether training which only recruits frontal lobe is retained over time. It may be the case that temporal lobe encoding is actually necessary for long-term learning of the speech contrast (Myers et al., under review). It is also unknown whether short-term learning in the frontal lobes reflects a different perceptual status of the stimulus as compared to when this sensitivity emerges in the temporal lobes. For instance, it is possible that frontal encoding relies more heavily on domain-general systems for perceptual categorization whereas temporal encoding reflects a more genuine status of the stimuli as phonetic categories.

A system which allowed for rapid, on-the-fly adaptation to new phonetic category structure might present several advantages not only for learning new speech contrasts, but also for processing details of native language speech. As listeners, we are exposed to speech variants that differ significantly from our native language phonetic categories, for instance, in the case of foreign accents, yet we are also able to quickly adapt to non-standard speech sounds (Bradlow and Bent, 2008; Kraljic et al., 2008). A neural system which likewise showed rapid, contextually-sensitive flexibility to shift phonetic category boundaries would facilitate this kind of adaptation. At the same time, unconstrained

flexibility in processing non-standard speech sounds could be disadvantageous—for instance, one's phonetic category boundaries should not be continuously perturbed by every exposure to a new talker or accent. As such, a separate neural system which shows more stable, slowly-adapting responses would be also advantageous.

Several testable predictions fall out of this type of model. First, if frontal-to-temporal feedback is necessary for non-native phonetic category learning, patients with frontal lobe pathology (e.g., individuals with Broca's aphasia) would have significant deficits in the acquisition and retention of new category information, while retaining sensitivities to native language phonetic category information learned pre-insult. Second, under the assumption that frontal systems are only engaged when category-level information is required for acquisition, it should be the case that incidental learning of phonetic categories, whether via sensitivity to statistical properties of the input (Hayes-Harb, 2007), or through other implicit methods (e.g., Lim and Holt, 2011; Vlahou et al., 2012) should be spared in this same population. Finally, if this frontal-to-temporal pathway is directed along the arcuate fasiculus, the coherence of this pathway should predict better speech sound learning at an individual level (see Myers et al., under review), and category training should be difficult for patients whose lesions implicate this pathway. Finally, as shown by Ventura-Campos et al. (2013), functional connectivity between frontal and posterior sites should inversely correlate with learning success as listeners transfer category-level learning to reshape perceptual sensitivities in the posterior temporal lobe.

CONCLUSION

The model described here is motivated largely through training studies which have used explicit, metalinguistic tasks in order to induce phonetic category sensitivities. There is still much to learn regarding phonetic category acquisition. First, little is known regarding the mechanisms which support encoding of statistical/distributional information which may reshape sensitivities "for free" as listeners are passively exposed to a new language. In the visual and auditory (non-speech) modalities, evidence suggests that medial temporal lobe and subcortical structures, in particular the caudate, may play a crucial role in encoding statistical regularities in the input (e.g., Turk-Browne et al., 2009; Durrant et al., 2013). Yet it is unknown whether the same structures mediate statistical learning for non-native speech sounds. At least one study (Golestani and Zatorre, 2004) showed engagement of the caudate for non-native speech sounds after training, although this result was attributed by these authors to the role of the caudate in motor speech control rather than in statistical learning.

Relatedly, the process of learning a non-native contrast involves encoding speech sounds in memory, but also protecting these newly-learned sounds from interference from existing similar speech sounds in one's native language. Recent work from our lab (Earle and Myers, under review) suggests that consolidation during sleep plays a significant role in this process. Participants who learned a non-native speech contrast in the evening showed improvements in discrimination of this contrast after an overnight interval and 24 h after learning, whereas participants who learned the same contrast in the morning did not show retention of the contrast after sleep. A follow-up suggested that the morning group's failure to retain the contrast was due to interference from exposure to similar native-language speech sounds over the course of the day. Taken together, this evidence suggests that (a) sleep plays a stabilizing role in the perceptual learning of speech sounds and (b) interference before sleep can serve to disrupt perceptual learning. This finding joins a literature on perceptual learning of synthetic speech sounds (Fenn et al., 2003, 2013) and on lexical learning which point to a crucial role for sleep in either abstracting away from the episodic details of the input, or to protection of learning from decay. While the neural bases of sleep-related consolidation for speech sounds have yet to be investigated, following a complementary systems memory framework (McClelland et al., 1995; O'Reilly and Rudy, 2001), one might predict that immediate encoding of novel speech sounds would implicate the hippocampus, while the overnight interval would serve to transfer this learning to cortical systems (e.g., Davis et al., 2009). This hippocampal-to-cortical transfer is thought to support abstraction from the episodic details of the signal to a more abstract representation of the input.

Perhaps most importantly it has yet to be determined whether second-language learning in immersion or in the classroom induces the same types of neural responses observed here. To fully understand the boundaries of plasticity in adult phonetic category learning, future research will need to be directed at these topics.

ACKNOWLEDGMENTS

This work was supported by NIH NIDCD grants R03 DC009495 and R01 DC013064 to Emily B. Myers, and NIH NICHD grant P01 HD001994 (Rueckl, PI). Thanks to Sayako Earle and Alexis Johns for helpful comments on an earlier version of this manuscript. The content is the responsibility of the author and does not necessarily represent official views of the NIH, NIDCD, or NICHD.

REFERENCES

Ahissar, M., and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* 8, 457–464. doi: 10.1016/j.tics.2004.08.011

Basso, A., Casati, G., and Vignolo, L. A. (1977). Phonemic identification defect in aphasia. *Cortex* 13, 85–95. doi: 10.1016/S0010-9452(77)80057-9

Belin, P., Zatorre, R. J., and Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Res. Cogn. Brain Res.* 13, 17–26. doi: 10.1016/S0926-6410(01)00084-2

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–312. doi: 10.1038/35002078

Best, C. C., and McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Lang. Speech* 46(pt 2–3), 183–216. doi: 10.1177/00238309030460020701

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: zulu click discrimination by English-speaking adults and infants. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 345–360. doi: 10.1037/0096-1523.14.3.345

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

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *J. Acoust. Soc. Am.* 101, 2299–2310. doi: 10.1121/1.418276

Burgaleta, M., Baus, C., Díaz, B., and Sebastián-Gallés, N. (2014). Brain structure is related to speech perception abilities in bilinguals. *Brain Struct. Funct.* 219, 1405–1416. doi: 10.1007/s00429-013-0576-9

Burnham, D. K., Earnshaw, L. J., and Clark, J. E. (1991). Development of categorical identification of native and non-native bilabial stops: infants, children and adults. *J. Child Lang.* 18, 231–260. doi: 10.1017/S0305000900011041

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., and Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage* 19, 113–124. doi: 10.1016/S1053-8119(03)00020-X

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

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

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex* 48, 882–887. doi: 10.1016/j.cortex.2011.05.017

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., and Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *J. Cogn. Neurosci.* 21, 803–820. doi: 10.1162/jocn.2009.21059

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

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

DeWitt, 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

Diaz, B., Baus, C., Escera, C., Costa, A., and Sebastian-Galles, N. (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16083–16088. doi: 10.1073/pnas.0805022105

Durrant, S. J., Cairney, S. A., and Lewis, P. A. (2013). Overnight consolidation aids the transfer of statistical knowledge from the medial temporal lobe to the striatum. *Cereb. Cortex* 23, 2467–2478. doi: 10.1093/cercor/bhs244

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science* 171, 303–306. doi: 10.1126/science.171.3968.303

Emberson, L. L., Liu, R., and Zevin, J. D. (2013). Is statistical learning constrained by lower level perceptual organization? *Cognition* 128, 82–102. doi: 10.1016/j.cognition.2012.12.006

Evans, S., Kyong, J. S., Rosen, S., Golestani, N., Warren, J. E., McGettigan, C., et al. (2013). The pathways for intelligible speech: multivariate and univariate perspectives. *Cereb. Cortex*. doi: 10.1093/cercor/bht083. [Epub ahead of print].

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., and Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition* 127, 427–438. doi: 10.1016/j.cognition.2013.02.007

Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition* 128, 280–286. doi: 10.1016/j.cognition.2013.04.007

Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature* 425, 614–616. doi: 10.1038/nature01951

Flege, J. E., MacKay, I. R., and Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *J. Acoust. Soc. Am.* 106, 2973. doi: 10.1121/1.428116

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316. doi: 10.1126/science.291.5502.312

Garell, P. C., Bakken, H., Greenlee, J. D. W., Volkov, I., Reale, R. A., Oya, H., et al. (2013). Functional connection between posterior superior temporal gyrus and ventrolateral prefrontal cortex in human. *Cereb. Cortex* 23, 2309–2321. doi: 10.1093/cercor/bhs220

Golestani, N., Molko, N., Dehaene, S., LeBihan, D., and Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cereb. Cortex* 17, 575–582. doi: 10.1093/cercor/bhk001

Golestani, N., and Pallier, C. (2007). Anatomical correlates of foreign speech sound production. *Cereb. Cortex* 17, 929. doi: 10.1093/cercor/bhl003

Golestani, N., Paus, T., and Zatorre, R. J. (2002). Anatomical correlates of learning novel speech sounds. *Neuron* 35, 997–1010. doi: 10.1016/S0896-6273(02)00862-0

Golestani, N., Price, C. J., and Scott, S. K. (2011). Born with an ear for dialects? Structural plasticity in the expert phonetician brain. *J. Neurosci.* 31, 4213–4220. doi: 10.1523/JNEUROSCI.3891-10.2011

Golestani, N., and Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Golestani, N., and Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain Lang.* 109, 55–67. doi: 10.1016/j.bandl.2008.01.005

Gow, D. W., Segawa, J. A., Ahlfors, S. P., and Lin, F.-H. (2008). Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates. *Neuroimage* 43, 614–623. doi: 10.1016/j.neuroimage.2008.07.027

Guenther, F. H., and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *J. Acoust. Soc. Am.* 100(2 pt 1), 1111–1121. doi: 10.1121/1.416296

Guenther, F. H., Husain, F. T., Cohen, M. A., and Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *J. Acoust. Soc. Am.* 106, 2900–2912. doi: 10.1121/1.428112

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., and Tourville, J. A. (2004). Representation of sound categories in auditory cortical maps. *J. Speech Lang. Hear. Res.* 47, 46–57. doi: 10.1044/1092-4388(2004/005)

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Lang. Res.* 23, 65–94. doi: 10.1177/0267658307071601

Hickok, G., Costanzo, M., Capasso, R., and Miceli, G. (2011). The role of Broca's area in speech perception: evidence from aphasia revisited. *Brain Lang.* 119, 214–220. doi: 10.1016/j.bandl.2011.08.001

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

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

Joanisse, M. F., Zevin, J. D., and McCandliss, B. D. (2007). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using FMRI and a short-interval habituation trial paradigm. *Cereb. Cortex* 17, 2084–2093. doi: 10.1093/cercor/bhl124

Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition* 107, 54–81. doi: 10.1016/j.cognition.2007.07.013

Krieger-Redwood, K., Gaskell, M. G., Lindsay, S., and Jefferies, E. (2013). The selective role of premotor cortex in speech perception: a contribution to phoneme judgements but not speech comprehension. *J. Cogn. Neurosci.* 25, 2179–2188. doi: 10.1162/jocn_a_00463

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3863–3868. doi: 10.1073/pnas.0600244103

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 979–1000. doi: 10.1098/rstb.2007.2154

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606–608. doi: 10.1126/science.1736364

Lee, Y. S., Turkeltaub, P., Granger, R., and Raizada, R. D. S. (2012). Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. *J. Neurosci.* 32, 3942–3948. doi: 10.1523/JNEUROSCI.3814-11.2012

Leech, R., Holt, L. L., Devlin, J. T., and Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J. Neurosci.* 29, 5234–5239. doi: 10.1523/JNEUROSCI.5758-08.2009

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417

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

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., and Desai, A. (2013). Neural dynamics of phonological processing in the dorsal auditory stream. *J. Neurosci.* 33, 15414–15424. doi: 10.1523/JNEUROSCI.1511-13.2013

Lim, S., and Holt, L. L. (2011). Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. *Cogn. Sci.* 35, 1390–1405. doi: 10.1111/j.1551-6709.2011.01192.x

Lisker, L., and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 384–422.

Lively, S., Logan, J., and Pisoni, D. (1993). Training Japanese listeners to indentify English /r/ and /l/. II: the role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* 94, 1242–1255. doi: 10.1121/1.408177

Logan, J., Lively, S., and Pisoni, D. (1991). Training Japanese listeners to identify English/r/ and /l/: a first report. *J. Acoust. Soc. Am.* 89, 874–886. doi: 10.1121/1.1894649

Maye, J., Weiss, D. J., and Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Dev. Sci.* 11, 122–134. doi: 10.1111/j.1467-7687.2007.00653.x

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, B101–B111. doi: 10.1016/S0010-0277(01)00157-3

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., and McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cogn. Affect. Behav. Neurosci.* 2, 89–108. doi: 10.3758/CABN.2.2.89

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Molholm, S., Mercier, M. R., Liebenthal, E., Schwartz, T. H., Ritter, W., Foxe, J. J., et al. (2014). Mapping phonemic processing zones along human perisylvian cortex: an electro-corticographic investigation. *Brain Struct. Funct*. 219, 1369–1383. doi: 10.1007/s00429-013-0574-y

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

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: an fMRI investigation. *Neuropsychologia* 45, 1463–1473. doi: 10.1016/j.neuropsychologia.2006.11.005

Myers, E. B., and Blumstein, S. E. (2008). The neural bases of the lexical effect: an fMRI investigation. *Cereb. Cortex* 18, 278. doi: 10.1093/cercor/bhm053

Myers, E. B., Blumstein, S. E., Walsh, E., and Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychol. Sci.* 20, 895–903. doi: 10.1111/j.1467-9280.2009.02380.x

Myers, E. B., and Swan, K. (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. *J. Cogn. Neurosci.* 24, 1695–1708. doi: 10.1162/jocn_a_00243

Naatanen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin. Neurophysiol.* 118, 2544–2590. doi: 10.1016/j.clinph.2007.04.026

Nahum, M., Nelken, I., and Ahissar, M. (2008). Low-level information and high-level perception: the case of speech in noise. *PLoS Biol.* 6:e126. doi: 10.1371/journal.pbio.0060126

Niziolek, C. A., and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J. Neurosci.* 33, 12090–12098. doi: 10.1523/JNEUROSCI.1008-13.2013

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., et al. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cereb. Cortex* 20, 2486–2495. doi: 10.1093/cercor/bhp318

O'Reilly, R. C., and Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol. Rev.* 108, 311–345. doi: 10.1037/0033-295X.108.2.311

Paavilainen, P., Mikkonen, M., Kilpeläinen, M., Lehtinen, R., Saarela, M., and Tapola, L. (2003). Evidence for the different additivity of the temporal and frontal generators of mismatch negativity: a human auditory event-related potential study. *Neurosci. Lett.* 349, 79–82. doi: 10.1016/S0304-3940(03)00787-0

Pallier, C., Bosch, L., and Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition* 64, B9–B17. doi: 10.1016/S0010-0277(97)00030-9

Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cogn. Sci.* 25, 711–731. doi: 10.1207/s15516709cog2505_5

Polka, L., Colantonio, C., and Sundara, M. (2001). A cross-language comparison of /d /–/ð/ perception: evidence for a new developmental pattern. *J. Acoust. Soc. Am.* 109, 2190–2201. doi: 10.1121/1.1362689

Raizada, R. D., and Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron* 56, 726–740. doi: 10.1016/j.neuron.2007.11.001

Raizada, R. D. S., Tsao, F. M., Liu, H. M., and Kuhl, P. K. (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cereb. Cortex* 20, 1–12. doi: 10.1093/cercor/bhp076

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

Rogalsky, C., Love, T., Driscoll, D., Anderson, S. W., and Hickok, G. (2011). Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system. *Neurocase* 17, 178–187. doi: 10.1080/13554794.2010.509318

Sehm, B., Schnitzler, T., Obleser, J., Groba, A., Ragert, P., Villringer, A., et al. (2013). Facilitation of inferior frontal cortex by transcranial direct current stimulation induces perceptual learning of severely degraded speech. *J. Neurosci.* 33, 15868–15878. doi: 10.1523/JNEUROSCI.5466-12.2013

Stein, M., Federspiel, A., Koenig, T., Wirth, M., Strik, W., Wiest, R., et al. (2012). Structural plasticity in the language system related to increased second language proficiency. *Cortex* 48, 458–465. doi: 10.1016/j.cortex.2010.10.007

Swan, K. S., and Myers, E. B. (2013). Category labels induce boundary-dependent perceptual warping in learned speech categories. *Second Lang. Res.* 29, 391–411. doi: 10.1177/0267658313491763

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., and Johnson, M. K. (2009). Neural evidence of statistical learning: efficient detection of visual regularities without awareness. *J. Cogn. Neurosci.* 21, 1934–1945. doi: 10.1162/jocn.2009.21131

Turkeltaub, P. E., and Branch Coslett, H. (2010). Localization of sublexical speech perception components. *Brain Lang.* 114, 1–15. doi: 10.1016/j.bandl.2010.03.008

Ventura-Campos, N., Sanjuán, A., González, J., Palomar-García, M.-Á., Rodríguez-Pujadas, A., Sebastián-Gallés, N., et al. (2013). Spontaneous brain activity predicts learning ability of foreign sounds. *J. Neurosci*. 33, 9295–9305. doi: 10.1523/JNEUROSCI.4655-12.2013

Vlahou, E. L., Protopapas, A., and Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *J. Exp. Psychol. Gen*. 141, 363–381. doi: 10.1037/a0025014

Wade, T., and Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J. Acoust. Soc. Am.* 118, 2618–2633. doi: 10.1121/1.2011156

Werker, J. F., and Tees, R. C. (1999). Influences on infant speech processing: toward a new synthesis. *Annu. Rev. Psychol.* 50, 509–535. doi: 10.1146/annurev.psych.50.1.509

Winkowski, D. E., Bandyopadhyay, S., Shamma, S. A., and Kanold, P. O. (2013). Frontal cortex activation causes rapid plasticity of auditory cortical processing. *J. Neurosci.* 33, 18134–18148. doi: 10.1523/JNEUROSCI.0180-13.2013

Wong, P. C. M., Perrachione, T. K., and Parrish, T. B. (2007). Neural characteristics of successful and less successful speech and word learning in adults. *Hum. Brain Mapp.* 28, 995–1006. doi: 10.1002/hbm.20330

Yeung, H. H., and Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition* 113, 234–243. doi: 10.1016/j.cognition.2009.08.010

Zatorre, R. J., Fields, R. D., and Johansen-Berg, H. (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nat. Neurosci.* 15, 528–536. doi: 10.1038/nn.3045

Zevin, J. D., Datta, H., Maurer, U., Rosania, K. A., and McCandliss, B. D. (2010). Native language experience influences the topography of the mismatch negativity to speech. *Front. Hum. Neurosci.* 4:212. doi: 10.3389/fnhum.2010.00212

Zhou, X., and Merzenich, M. M. (2007). Intensive training in adults refines A1 representations degraded in an early postnatal critical period. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15935–15940. doi: 10.1073/pnas.0707348104

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: 24 March 2014; accepted: 20 July 2014; published online: 08 August 2014.
Citation: Myers EB (2014) Emergence of category-level sensitivities in non-native speech sound learning. Front. Neurosci. ***8****:238. doi: 10.3389/fnins.2014.00238*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 03 September 2014
doi: 10.3389/fnins.2014.00275

Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners

***Daniel Callan*[1,2]*, *Akiko Callan*[1,2] *and Jeffery A. Jones*[3]**

[1] *Center for Information and Neural Networks, National Institute of Information and Communications Technology, Osaka University, Osaka, Japan*
[2] *Multisensory Cognition and Computation Laboratory Universal Communication Research Institute, National Institute of Information and Communications Technology, Kyoto, Japan*
[3] *Laurier Centre for Cognitive Neuroscience and Department of Psychology, Wilfrid Laurier University, Waterloo, ON, Canada*

Edited by:
Lynne E. Bernstein, George Washington University, USA

Reviewed by:
Matthew H. Davis, MRC Cognition and Brain Sciences Unit, UK
Neil M. McLachlan, The University of Melbourne, Australia

***Correspondence:**
Daniel Callan, Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Osaka University, 1-4 Yamadaoka, Suita City, Osaka 565-0871, Japan
e-mail: dcallan@nict.go.jp

Brain imaging studies indicate that speech motor areas are recruited for auditory speech perception, especially when intelligibility is low due to environmental noise or when speech is accented. The purpose of the present study was to determine the relative contribution of brain regions to the processing of speech containing phonetic categories from one's own language, speech with accented samples of one's native phonetic categories, and speech with unfamiliar phonetic categories. To that end, native English and Japanese speakers identified the speech sounds /r/ and /l/ that were produced by native English speakers (unaccented) and Japanese speakers (foreign-accented) while functional magnetic resonance imaging measured their brain activity. For native English speakers, the Japanese accented speech was more difficult to categorize than the unaccented English speech. In contrast, Japanese speakers have difficulty distinguishing between /r/ and /l/, so both the Japanese accented and English unaccented speech were difficult to categorize. Brain regions involved with listening to foreign-accented productions of a first language included primarily the right cerebellum, left ventral inferior premotor cortex PMvi, and Broca's area. Brain regions most involved with listening to a second-language phonetic contrast (foreign-accented and unaccented productions) also included the left PMvi and the right cerebellum. Additionally, increased activity was observed in the right PMvi, the left and right ventral superior premotor cortex PMvs, and the left cerebellum. These results support a role for speech motor regions during the perception of foreign-accented native speech and for perception of difficult second-language phonetic contrasts.

Keywords: speech perception, accent, fMRI, Broca's area, premotor, cerebellum, internal model, non-native speech

INTRODUCTION

A growing body of research suggests that speech motor areas are recruited to facilitate auditory speech perception when the acoustic signal is degraded or masked by noise (Callan et al., 2010; Schwartz et al., 2012; Adank et al., 2013; Moulin-Frier and Arbib, 2013). Researchers hypothesize that auditory speech signals are translated into internally simulated articulatory control signals (articulatory-auditory internal models), and that these internal simulations help to constrain speech perception (Callan et al., 2004a; Wilson and Iacoboni, 2006; Skipper et al., 2007; Iacoboni, 2008; Poeppel et al., 2008; Rauschecker, 2011; Schwartz et al., 2012). Indeed, brain imaging studies have demonstrated that activity increases in speech motor areas when participants listen to speech in noise relative to when they listen in noise-free conditions (Callan et al., 2003a, 2004b). Increased activity in speech motor areas has also been observed when listeners identify phonetic categories that are not in their first language (non-native), relative to the activity observed when they identify phonetic categories from their first language (native) (Callan et al., 2003b, 2004a, 2006a; Wang et al., 2003). Moreover, activity in speech motor areas has been found to increase when participants listen to sentences in their first language when they are spoken in an unfamiliar accent (Adank et al., 2013). These observations, as well as observations from other studies that have demonstrated that speech motor brain regions are responsive to both production and perception of speech, support motor simulation theories of speech perception (Callan et al., 2000, 2006b, 2010; Wilson et al., 2004; Nishitani et al., 2005; Meister et al., 2007). In this study, we investigated the neural processes involved in the perception of phonetic categories from one's first language produced by native speakers, as well as those produced by speakers with a foreign-language accent. We compared the neural activity in these conditions to the activity observed when participants perceived phonetic categories from their second language (again, both produced by a native speaker of that second language, and produced by a speaker with a foreign-language accent).

Adults often have considerable difficulty discriminating and identifying many non-native phonetic categories in their second

language that overlap with a single phonetic category in their first (native) language, even after years of exposure to that second language (Miyawaki et al., 1975; Trehub, 1976; Strange and Jenkins, 1978; Werker et al., 1981; Werker and Tees, 1999). The English /r/ and /l/ phonetic contrast is an example of a difficult non-native phonetic contrast for native Japanese speakers (Miyawaki et al., 1975). Intensive phonetic identification training can result in long-term improvement in speech perception that generalizes to novel stimuli (Lively et al., 1994; Akahane-Yamada, 1996; Bradlow et al., 1999). Perceptual identification training can also lead to improvements in production (Bradlow et al., 1997), even in the absence of formal production training. The observation that perceptual improvements lead to production improvements suggests that a perceptual-motor component may be responsible for the improved phonetic identification. Indeed, several brain-imaging studies support the hypothesis that neural processes associated with speech production constrain and facilitate phoneme identification (Callan et al., 2004a, 2010; Skipper et al., 2007).

Similar to the difficulties listeners have discriminating and identifying non-native phonetic contrasts in a second language, foreign-accented native speech is often difficult for a native speaker of the language to perceive (Goslin et al., 2012; Adank et al., 2013; Moulin-Frier and Arbib, 2013). Recent evidence suggests that speech motor processes are recruited to facilitate perception when listening to foreign-accented productions of a language (Adank et al., 2013; Moulin-Frier and Arbib, 2013). For example, Adank et al. (2013) found evidence for sensorimotor integration during processing of foreign-accented speech when they asked one group of participants to imitate the unfamiliar foreign-accent of a speaker who uttered sentences in the participants' first language, and compared their brain activity to another group of participants who repeated the same sentences in their own native accent. Adank et al. (2013) compared the levels of activation in the speech motor regions of the brain (including the inferior frontal gyrus, and Broca's area) when participants listened to sentences before a production task, to the levels of activation observed when participants listened to sentences after a production task. Larger differences in speech motor activity were observed for the participants who imitated the unfamiliar, foreign-accented speech, compared to the participants who repeated the sentences in their own accent, specifically when the participants listened to the sentences before compared to after the production task.

The goal of the present study was to differentiate the neural processes that are involved in the perception of phonetic categories in a second language (non-native), from the neural processes involved in the perception of foreign-accented productions of phonetic categories from one's first language. In this study, native English (Eng) and Japanese (Jpn) speakers listened to native English ("unaccented") and Japanese ("accented") productions of English syllables that began with either /r/ or /l/. The Japanese productions of the English syllables (accented) used for the study were found to have a confusion rate (misidentified as the wrong syllable) of 29% when presented to native English speakers. The Japanese-accented productions could be perceived as either /r/ or /l/ by native English speakers on a proportion of the trials. The native English speakers were more accurate at identifying the unaccented English speech stimuli than the Japanese-accented speech stimuli. In contrast, the native Japanese speakers had difficulty identifying both the English-unaccented speech stimuli and the Japanese-accented stimuli. The following contrasts were investigated: (1) The neural processes that are involved in the perception of foreign-accented productions of a first language phonetic category were investigated using the contrast Eng(accented – unaccented) – Jpn(accented – unaccented). Subtracting the activity observed in the Jpn group controlled for general stimulus variables. (2) The contrast of Eng(accented) – Eng(unaccented) investigated which areas were involved in processing a difficult native phonetic identification task (accented) compared to those involved in processing an easy phonetic identification task (unaccented), without the potential confound of extraneous between group differences. However, acoustic stimulus characteristics were not controlled for by this contrast. (3) The neural processes selective for the perception of foreign-accented productions of a second language phonetic category, compared to foreign-accented productions of a first language phonetic category, were investigated using the contrast Jpn(accented) – Eng(accented). This contrast controlled for the neural processes that were related to task difficulty, such as attention and verbal rehearsal. (4) To investigate the overall neural processes involved in the perception of (native) unaccented productions of a second language phonetic category relative to the perception of unaccented productions of a first language phonetic category, we used the contrast Jpn(unaccented) – Eng(unaccented). This contrast did not control for task difficulty. All three of the contrasts above controlled for general processes related to performing a categorical perceptual identification task using a button response, though only the Jpn(accented) – Eng(accented) contrast additionally controlled for task difficulty.

A number of brain regions have been shown to be involved with the perception of unaccented/native productions of a second language phonetic category (Callan et al., 2003a, 2004a, 2006a; Wang et al., 2003) as well as foreign-accented speech (Adank et al., 2013). These regions include, but are not limited to: the ventral inferior premotor cortex including Broca's area (PMvi), the ventral superior and dorsal premotor cortex (PMvs/PMd), the superior temporal gyrus/sulcus (STG/S), and the cerebellum. If the neural processes involved in processing difficult-to-perceive speech sounds are dependent on the relative contribution of regions involved in articulatory planning control, then one might predict that the brain regions involved with speech motor control (PMvi/Broca's, PMvs/PMd, and the cerebellum) would be more active than regions involved with auditory processing (STG/S) when general acoustic differences in the stimuli are controlled.

As previously mentioned, the brain regions involved with internally simulating speech production (internal models) are hypothesized to constrain and facilitate speech perception, especially under degraded conditions (e.g., speech in noise, non-native speech) (Callan et al., 2003b, 2004a; Iacoboni and Wilson, 2006; Wilson and Iacoboni, 2006; Skipper et al., 2007; Iacoboni, 2008; Rauschecker and Scott, 2009; Rauschecker, 2011; Callan et al., 2014). Internal models are thought to simulate the input/output characteristics, or their inverses, of the motor

control system (Kawato, 1999). With regards to speech production, inverse internal models predict the motor commands necessary to articulate a desired auditory (and/or orosensory) target (auditory-to-articulatory mapping). Forward internal models, conversely, predict the auditory (and/or orosensory) consequences of simulated speech articulation (articulatory-to-auditory mapping). It has been proposed that both forward and inverse internal models constrain and facilitate speech perception, especially under degraded conditions (Callan et al., 2004a, 2014; Rauschecker and Scott, 2009; Rauschecker, 2011). Facilitation is achieved by a process akin to analysis-by-synthesis (Stevens, 2002; Poeppel et al., 2008) (forward internal models: articulatory-to-auditory prediction) and synthesis-by-analysis (inverse internal models: auditory-to-articulatory prediction), specifically by competitive selection of the speech unit (phoneme, syllable, etc.) that best matches the ongoing auditory signal (or visual signal, in the case of audiovisual or visual-only speech). Brain regions thought to be involved with instantiating these articulatory-to-auditory and auditory-to-articulatory internal models include speech motor areas such as the PMC and Broca's area, the posterior regions of the STG/S, the IPL, and the cerebellum. In particular, the cerebellum, has been shown to instantiate internal models for motor control (Kawato, 1999; Imamizu et al., 2000), and there is evidence that it instantiates internal models related to speech (Callan et al., 2004a, 2007; Rauschecker, 2011; Tourville and Guenther, 2011; Callan and Manto, 2013). Brain activity in these regions (including the PMC, Broca's area, the IPL, and the cerebellum) during speech perception tasks has been used as evidence to support the involvement of motor processes during speech perception.

One potential criticism of ascribing activity found in speech motor regions to speech perception is that many of these same regions are known to be more active as a function of task difficulty. Activity in brain regions such as the IFG, the PMC, and the cerebellum has been shown to increase with task-related attentional demands and working memory (including verbal rehearsal) (Jonides et al., 1998; Davachi et al., 2001; Sato et al., 2009; Alho et al., 2012). As has been previously suggested (Hickok and Poeppel, 2007; Poeppel et al., 2008; Lotto et al., 2009; Scott et al., 2009), activity in these speech motor regions may not be related to speech perception intelligibility, but rather to other processes related to task difficulty. If these brain regions involved with speech motor processing are increasingly more active as a function of task difficulty, one would predict that subjects with worse phonetic identification performance (greater task difficulty) would show increased activity in these regions compared to subjects with better phonetic identification performance. However, the opposite result has been found, with an increase in PMC, IFG, and cerebellum activity associated with better phonetic identification performance on a difficult non-native phonetic category (Callan et al., 2004a). Similarly, PMC activity has been shown to be more active for correct compared to incorrect trials during a phonetic identification in noise task (Callan et al., 2010).

It is hypothesized that the perception of foreign-accented first language phonetic categories depends on the brain regions that instantiate the auditory—articulatory representation of phonetic categories. Research suggests that these regions include left hemisphere Broca's area and the PMC. In the case of the perception of second-language phonetic categories—for which the distinct second-language phonemes are subsumed within a single phonetic category in the native language (e.g., English /r/ and /l/ for native Japanese speakers)—additional neural processes may be recruited to establish new phonetic categories without interfering with the established native phonetic category. It is hypothesized that the establishment of these second-language phonetic categories (when the second-language is acquired after childhood) involves greater reliance on general articulatory-to-auditory feedback control systems, which generate auditory predictions based on articulatory planning, and are thought to be instantiated in right hemisphere PMC (Tourville and Guenther, 2011; Guenther and Vladusich, 2012).

METHODS

SUBJECTS

Thirteen right-handed native Japanese (Jpn) speakers with some English experience (at least 6 years of classes in junior and senior high school) and thirteen right-handed native English (Eng) speakers participated in this study. The native Japanese-speaking subjects were nine females and four males whose ages ranged from 23 to 37 years ($M = 30.4$ years, $SD = 4.5$). The native English-speaking subjects were one female and twelve males whose ages ranged from 21 to 39 years ($M = 27.8$ years, $SD = 5.1$). All subjects included in this study scored significantly above chance when they identified the /r/ and /l/ productions of a native English speaker, which ensured that all subjects were actively trying to do the task. Subjects were paid for their participation, and gave written informed consent for the experimental procedures, which were approved by the ATR Human Subject Review Committee in accordance with the principles expressed in the Declaration of Helsinki.

STIMULI AND PROCEDURE

The stimuli were acquired from the speech database compiled by the Department of Multilingual Learning (ATR—HIS, Kyoto, Japan). The experiment had two, within-subject conditions: a foreign-accented speech condition and an unaccented speech condition. These two conditions were composed of audio speech stimuli consisting of English syllables beginning with a /r/ or /l/, which were followed by five different following English vowel contexts (/a, e, i, o, u/). There were three occurrences of each syllable for each accent condition for a total of 60 trials in the experiment. All stimuli were recorded digitally in an anechoic chamber with a sampling rate of 44,100 Hz. The unaccented speech was taken from samples of female and male native English speakers. The foreign-accented speech was taken from samples of female and male native Japanese speakers that produced /r/–/l/ confusions ($M = 29\%$, $SD = 13\%$), as determined by a forced-choice identification task performed by native English speakers (the number of evaluators ranged from 6 to 10 individuals, depending on the stimulus). Both the foreign-accented and unaccented /r/ and /l/ stimuli consisted of six female voices and nine male voices. The stimuli were down-sampled to 22,050 Hz for presentation during the experiment.

The fMRI procedure consisted of an event-related design in which the sequence of presentation of the various stimulus conditions (unaccented /r/, unaccented /l/, foreign-accented /r/, foreign-accented /l/, and /null trial/) was generated stochastically using SPM99 (Wellcome Department of Cognitive Neurology, UCL). An event-related design was employed so that the various stimulus conditions could be presented (approximately 85–90 dB SPL) in a pseudo-random order. This ensured that subjects could not predict which stimulus would occur during the subsequent presentation. Stimuli were presented (synchronized with fMRI scanning using Neurobehavioral System's Presentation software) via MR-compatible headphones (Hitachi Advanced Systems' ceramic transducer headphones; frequency range 30–40,000 Hz, approximately 20 dB SPL passive attenuation). Subjects identified whether the stimuli started with /r/ or /l/, and indicated which they perceived by pressing a button with their left thumb. The left hand was used instead of the right hand so that brain activity in left Broca's area and left PMC could be better identified, with less influence of activity associated with the button-press motor response. The identity of the buttons was counterbalanced across subjects. Stimuli were presented at a rate of approximately 2250 ms in a pseudo-random order dependent on the event sequence. Subjects were asked to respond quickly to minimize differences in the hemodynamic response resulting from long response times (Poldrack, 2000). However, they were not asked to respond as quickly as they could, therefore response latencies were not evaluated. Null trials in which only silence occurred were also included and used as a baseline condition. Subjects were not given online feedback regarding the correctness of their responses. All subjects were given a practice session outside of the scanner using stimuli similar to those used in the experimental session.

Each subject participated in multiple experiments, including the present study, within the same insertion into the fMRI scanner. The order of the different experiments was counterbalanced across subjects. Depending on the number of experiments in which a subject participated, the total time in the scanner ranged from approximately 30–60 min. The session lasted approximately 7 min for this experiment.

fMRI DATA COLLECTION AND PREPROCESSING

For functional brain imaging, Shimadzu-Marconi's Magnex Eclipse 1.5T PD250 was used at the ATR Brain Activity Imaging Center. Functional T2* weighted images were acquired using a gradient echo-planar imaging sequence (echo time 55 ms; repetition time 2000 ms; flip angle 90°). A total of 20 contiguous axial slices were acquired with a $3 \times 3 \times 6$ mm voxel resolution covering the cortex and cerebellum. For some subjects, 20 slices was not a sufficient number to cover the entire cortex and thus the top part of the cortex was missing. As a result, the analyses conducted in this study do not include the top part of the cortex. A total of 304 scans were taken during a single session. Images were preprocessed using programs within SPM8 (Wellcome Department of Cognitive Neurology, UCL). Differences in acquisition time between slices were accounted for; images were realigned and spatially normalized to a standard space using a template EPI image ($3 \times 3 \times 3$ mm voxels), and were smoothed using a $6 \times 6 \times 12$ mm FWHM Gaussian kernel.

STATISTICAL IMAGE ANALYSIS

Regional brain activity for the various conditions was assessed with a general linear model using an event-related design. Realignment parameters were used to regress out movement-related artifacts. In addition, low-pass filtering, which used the hemodynamic response function, was employed. The event-related stochastic design used to model the data included null responses and a stationary trial occurrence probability. A mixed-effects model was employed. A fixed-effect analysis was first employed for all contrasts of interest across data from each subject separately. The contrasts of interest for both the Jpn and Eng subjects included: unaccented speech relative to baseline; accented speech relative to baseline; and accented relative to unaccented speech. At the random effects level between subjects, the contrast image of the parameter estimates of the first level analysis for each subject was used as input for a SPM model employing two-sample *t*-tests. The contrasts of interest consisted of the following: (1) Processes related to the perception of first language phonetic contrasts in accented speech Eng(accented – unaccented) – Jpn(accented – unaccented); (2) Processes related to the perception of first language accented speech (difficult task) relative to first language unaccented speech (easy task). (3) Processes related to the perception of foreign-accented speech Jpn(accented) – Eng(accented) and (4) Processes related to the perception of unaccented productions of a second language phonetic category Jpn(unaccented) – Eng(unaccented). Because the study is quasi-experimental in the sense that assignment of participant into Eng and Jpn groups is not random, the variance not attributable to the independent experimental variables (e.g., educational experience and cultural differences related to carrying out the tasks) may significantly influence participants' performance and neural responses, which could potentially confound the results. To ensure that the differential brain activity related to the contrasts of interest (given above) were not the result extraneous neural processes involved with behavioral performance, task difficulty, and/or variables arising from the quasi-experimental design, the random-effects analyses were conducted using the raw percent correct phonetic identification performance scores as a covariate of non-interest.

A False Discovery Rate (FDR) correction for multiple comparisons across the entire volume was employed with a threshold of $pFDR < 0.05$ using a spatial extent greater than 5 voxels. If no voxels were found to be significant using the FDR, a correction threshold of $p < 0.001$ uncorrected with a spatial extent threshold greater than 5 voxels was used. Region of interest (ROI) analyses were conducted using MNI coordinates for the PMvi/IFG (left −51,9,21; right 51,15,18), the PMvs (left −36,−3,57; right 27,−3,51), the STG/S (left −57,−39,9) and the cerebellum (left −27,−63,−39; right 30,−66,−33) given that in Callan et al. (2004a) these regions were found to be involved in processing difficult-to-perceive speech contrasts. It should be noted that these coordinates (for PMvi/IFG and STG/S) fall within the cluster of activity in regions found to be active for perception of accented speech, as reported by Adank et al. (2013). Small volume

correction for multiple comparisons was carried out using the seed voxels reported above within a sphere with a radius of 8 mm. The location of active voxels was determined by reference to the Talairach atlas (Talairach and Tournoux, 1988) as well as by using the Anatomy Toolbox within SPM8. Activity in the cerebellum was localized with reference to the atlas given by Schmahmann et al. (2000).

RESULTS

BEHAVIORAL PERFORMANCE

The results of the two-alternative forced-choice phoneme identification task (in percent correct) were analyzed across subjects using an ANOVA with the two factors of language group (Jpn and Eng) and accent (unaccented and accented). Bonferroni corrections for multiple comparisons were used to determine statistical significance at $p < 0.05$ for all behavioral analyses conducted. The results are as follows: the interaction between Jpn and Eng subjects for accented and unaccented stimuli was significant [Eng unaccented: $M = 94.6\%$, $SE = 0.6$; Jpn unaccented: $M = 69.5\%$, $SE = 1.8$; Eng accented: $M = 65.0\%$, $SE = 1.5$; Jpn accented: $M = 62.5\%$, $SE = 2.8$; $F_{(1,\ 48)} = 40.2$, $p < 0.05$ corrected] (see **Figure 1**). The main effect of group (Eng > Jpn) was significant [Eng $M = 79.8\%$, $SE = 3.11$, Jpn $M = 66.0\%$, $SE = 1.8$, $F_{(1,\ 48)} = 60.3$, $p < 0.05$ corrected]. The main effect of accent (unaccented > accented) was also significant [unaccented: $M = 82.1\%$, $SE = 2.9$, accented: $M = 63.7\%$, $SE = 1.1$, $F_{(1,\ 48)} = 106.4$, $p < 0.05$ corrected]. The identification performance on the two-alternative forced-choice task was significantly greater than chance for the unaccented and accented conditions for both Eng and Jpn subjects (see **Figure 1**) [Jpn unaccented: $T_{(12)} = 7.2$, $p < 0.05$ corrected; Jpn accented: $T_{(12)} = 7.2$, $p < 0.05$ corrected; Eng unaccented: $T_{(12)} = 75.1$, $p < 0.05$ corrected; Eng accented: $T_{(12)} = 10.7$, $p < 0.05$ corrected]. The Eng subjects had significantly better performance than the Jpn subjects for the unaccented speech stimuli condition [$T_{(12)} = 9.4$; $p < 0.05$ corrected]. For accented stimuli, there was no significant difference for identification (evaluated based on the intended production of the stimuli) between native English speaking subjects and native Japanese speaking subjects [$T_{(24)} = 1.1$; $p = 0.27$ uncorrected]. There was also no significant difference between Eng subjects' performance for the accented stimuli and Jpn subjects' performance for the unaccented stimuli [$T_{(24)} = 1.13$, $p = 0.15$ uncorrected]. For Eng subjects there was a significant difference between performance for the unaccented and accented stimuli [$T_{(12)} = 18.2$, $p < 0.05$ corrected]. The difference for Jpn subjects between the performance for unaccented and accented stimuli was not significant when corrections were made for multiple comparisons, but the difference was significant using an uncorrected threshold [$T_{(12)} = 3.3$, $p < 0.01$ uncorrected].

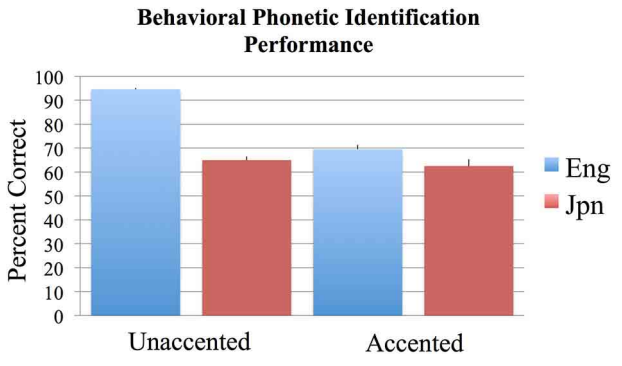


FIGURE 1 | Mean percent correct behavioral phonetic (/r/ vs. /l/) identification performance for the English (blue) and Japanese (red) groups for unaccented and foreign-accented speech. Standard error of the mean is given above each bar. All conditions were significantly above chance performance of 50%. See text for additional contrasts that were statistically significant.

BRAIN IMAGING

The random effects one-sample *t*-test of the unaccented and accented condition relative to the null condition (background scanner noise) was carried out separately for Jpn and Eng groups. A FDR correction for multiple comparisons across the entire volume was used with a threshold of pFDR < 0.05 (spatial extent > 5 voxels). The results for unaccented and accented conditions for both the Eng and Jpn groups (see **Figures 2A–D**) indicated extensive activity in regions of the brain known to be involved with speech processing bilaterally (STG/S, including primary auditory cortex, MTG, SMG, Broca's area, PMC, medial frontal cortex MFC/pre-suplementary motor area pre-SMA, anterior cingulate cortex ACC, cerebellar lobule VI, cerebellar Crus I). Activity associated with the motor response of pushing the button with the left thumb was also present for both the Jpn and Eng groups in the right motor and somatosensory cortex. The conjunction analysis, which determined the intersection of active voxels for all conditions thresholded at pFDR < 0.05, showed activity in most of the above-mentioned regions (see **Figure 2G** and **Table 1**).

The interaction effect between the factors of language group and accent is discussed below. The main effect of accent (accented vs. unaccented) did not show any significant differential activity using a corrected threshold of pFDR < 0.05 or an uncorrected threshold of $p < 0.001$ (spatial extent > 5 voxels). The main effect of language group (Jpn vs. Eng, see **Figure 2H** and **Table 2**) showed significant differential activity for Japanese > English (red) $p < 0.001$ (spatial extent > 5 voxels), predominantly in left and right PMvi/Broca's area, PMvs/PMd, the postcentral gyrus, the cerebellum, and the left inferior parietal lobule. The significant differential activity for Eng > Jpn (blue) $p < 0.001$) (spatial extent > 5 voxels) was present predominantly in the medial frontal gyrus, the middle frontal gyrus, the anterior cingulate cortex, and the middle cingulate cortex.

The contrast of accented relative to unaccented speech was carried out separately for Eng and Jpn subjects. For both Eng and Jpn subjects, no significant activity was found using a corrected threshold of pFDR < 0.05; therefore, a threshold of $p < 0.001$ uncorrected was used. For Eng subjects, activity was found to be present in left PMvi/Broca's area, right PMvs/PMD, left Broca's area BA 45, left IFG BA 47, the pre-SMA, and left and right cerebellar lobules VI and VIIa (see **Figure 2E** and **Table 3**). The results of the region of interest analysis (ROI) using small volume correction for multiple comparisons revealed significant activity in the left and right cerebellum lobule VI, and a trend toward

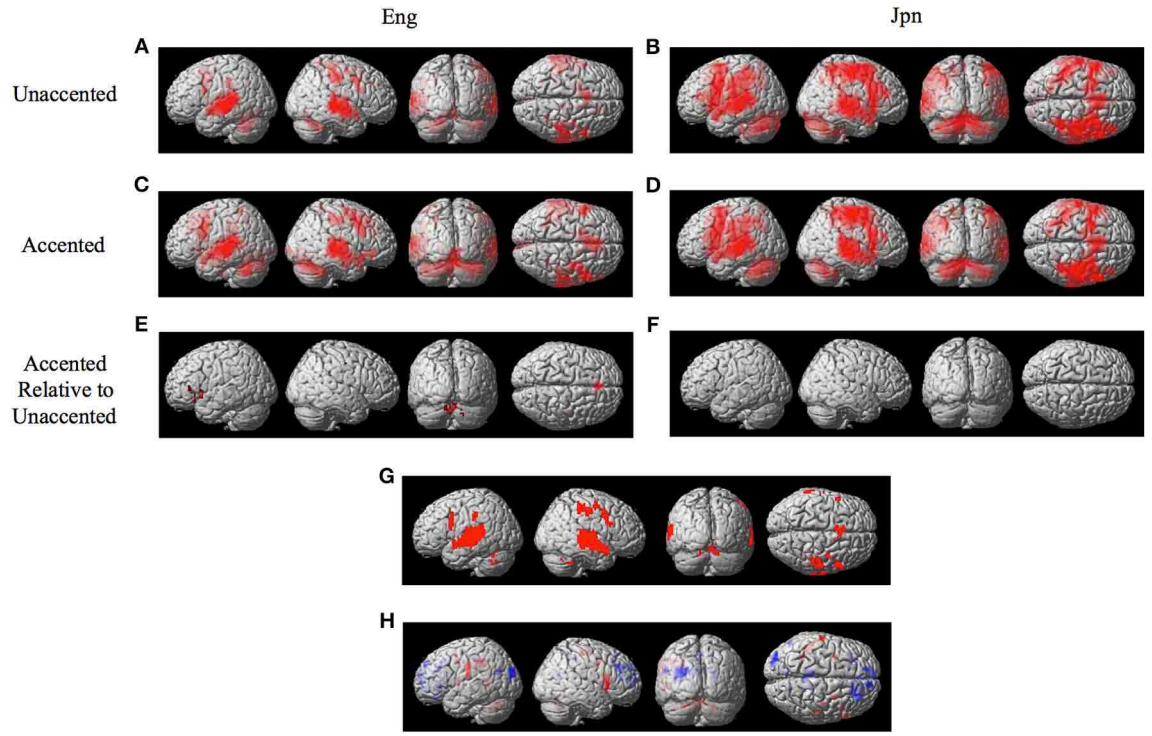


FIGURE 2 | Significant brain activity (thresholded at pFDR < 0.05 corrected) for the contrast of (A) Eng (unaccented), (B) Jpn (unaccented), (C) Eng (accented), and (D) Jpn (accented). All contrasts showed activity bilaterally in premotor cortex and Broca's area, the superior temporal gyrus/sulcus, the inferior parietal lobule, the pre-supplementary motor area pre-SMA, and the cerebellum. The conjunction analysis, shown in **(G)**, confirmed these regions were active for all conditions **(E)**. The contrast of accented Relative to unaccented thresholded at $p < 0.001$ uncorrected for Jpn showed activity in the left inferior frontal gyrus in Broca's area 44, the right dorsal premotor cortex, the pre-SMA, and the cerebellum bilaterally **(F)**. The contrast of accented relative to unaccented for the Jpn group did not show any significant activity thresholded at $p < 0.001$ uncorrected. The main effect of language group (Japanese vs. English) is shown in **(H)**, red corresponds to activity thresholded at $p < 0.001$ for Japanese > English and blue corresponds to activity for English > Japanese.

significant activity in the left PMvi/Broca's, the right PMvs/PMd, and the left STG/S (see **Table 4**). To ensure that the differential brain activity reported in the analyses of this study was not just the result extraneous neural processes involved with (or resulting from) behavioral performance, task difficulty (e.g., attention, working memory, concentration and/or response confidence), and/or variables arising from the quasi-experimental design, the same analyses were conducted using phonetic identification performance as a covariate of non-interest. The results of the contrast Eng(accented) – Eng(unaccented) using phonetic identification performance as a covariate of non interest showed activity in left PMvi/Broca's area, left Broca's BA 45, pre-SMA, right cerebellum Lobule VI, and left cerebellum lobule VII (see **Table 3**). The ROI analysis using phonetic identification performance as a covariate of non-interest revealed significant activity in left and right cerebellum lobule VI, and a trend toward significant activity in left PMvs ($p < 0.057$) (see **Table 4**). No significant activity was found for Jpn subjects using a threshold of $p < 0.001$ uncorrected or for the ROI analyses (see **Figure 2F** and **Tables 3**, **4**).

In order to determine brain activity that was related to difficult perceptual identification of a native phonetic contrast, the foreign-accented condition (which was difficult to perceive for both the native English speakers and the native Japanese speakers) was compared to the unaccented condition (which was easy to perceive for the native English speakers, but more difficult to perceive for the native Japanese speakers) between the Eng vs. the Jpn group using the contrast Eng(accented – unaccented) – Jpn(accented – unaccented) (random effects two-sample *t*-test). Only the pre-SMA activity was significant at $p < 0.05$ FDR corrected, therefore the analysis was conducted using a threshold of $p < 0.001$ uncorrected. Brain regions that showed significant differential activity for this contrast included the left and right Broca's area BA45, the pre-SMA, the right dorsolateral prefrontal cortex (DLPFC), the cerebellum lobule VIIa, and the brain stem (see **Figure 3** and **Table 3**). The same analysis using phonetic identification performance as a covariate of non-interest revealed activity only in left Broca's area using a threshold of $p < 0.0015$. The results of the ROI analysis using small volume correction for multiple comparisons revealed significant activity in the left PMvi, and the right cerebellum lobule VI (see **Figure 4** and **Table 4**). When using performance as a covariate of non-interest, no significant differential activity was found when correcting for multiple comparisons within the ROIs (**Table 4**).

Brain activity related to processing of foreign-accented productions of a second language phonetic category that was different from processing of foreign-accented productions of a first

Table 1 | Conjunction of all conditions Eng Unaccented, Jpn Unaccented, Eng Accented, Jpn Accented (Figure 2G).

Brain region	MNI coordinates
PMvi, Broca's area, BA 6,44	−54,12,27
	51,6,21
PMvs/PMd BA 6	−51,6,39
	54,9,39
	39,−12, 51
PostCG, IPL BA1,2	54,−30, 51
	−45,−30,39
Medial Frontal Cortex BA 9 Pre-SMA	−6,12,57
SPL BA7	−27,−57,45
Insula BA13	−36,−33,24
MTG/STG BA21,22	−63,−27,−3
	66,−27,−3
Cerebellum Vermis	0,−78,−18
Cerebellum Lobule VI	−18,−54,−24
	27,−66,−27

Table showing clusters of activity for the conjunction of all contrasts relative to rest (Eng Unaccented, Eng Unaccented, Jpn Accented, Jpn Unaccented) thresholded at pFDR < 0.05 corrected with an extent threshold greater than 5 voxels. Jpn., Japanese; Eng., English; Cor., corrected for multiple comparisons; BA, Brodmann area; PMvi, Ventral inferior premotor cortex; PMvs, Ventral superior premotor cortex; PMd, Dorsal premotor cortex; PostCG, Postcentral gyrus; IPL, Inferior parietal lobule; pre-SMA, Pre-supplementary motor area; SPL, Superior parietal lobule; MTG, Middle temporal gyrus; STG, Superior temporal gyrus. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity.

language phonetic category was investigated using the contrast Jpn(accented) – Eng(accented). No significant activity was found using a corrected threshold of pFDR <0.05, therefore a threshold of $p < 0.001$ uncorrected was used. Activity was present in the right PMvi/Broca's area and the right PMvs/PMd. Using phonetic identification performance as a covariate of non-interest revealed activity in right PMvi/Broca's area and right PMvs/PMd (see **Figure 5A** and **Table 3**). Using phonetic identification performance as a covariate of non-interest revealed activity in right PMvi/Broca's area, the right PMvs/PMd, and the left cerebellar lobule VI. For the ROI analysis, activity was significant in the right PMvi/Broca's area, right PMvs/PMd, and the left cerebellar lobule VI (see **Figure 6** and **Table 4**). Using performance as a covariate of non-interest, the ROI analysis showed significant activity in left cerebellar lobule VI, and a trend toward significance in both right PMvi/Broca's area ($p < 0.074$) and right PMvs/PMd ($p < 0.063$) (see **Table 4**).

To determine activity related to processing of unaccented productions of a second language phonetic category that was different from that of unaccented productions of a first language phonetic category, the difference between the Jpn and Eng subjects for unaccented speech was investigated using the contrast Jpn(unaccented) – Eng(unaccented). No significant activity was found using a corrected threshold of pFDR < 0.05, therefore, a threshold of $p < 0.001$ uncorrected was used. Activity was present in left and right PMvi/Broca's area, right PMvs/PMd, right Boca's BA45, left IFG BA47, left PostCG, left IPL, and left cerebellar lobules VIIa and V, as well as left and right cerebellar lobule VI (see **Figure 5B** and **Table 3**). Using phonetic identification performance as a covariate of non-interest, the analysis revealed activity primarily in right PMvi/Broca's area, right PMvs/PMd, and left cerebellum lobule VI. The results of the ROI analysis using small volume correction for multiple comparisons revealed significant

Table 2 | Main contrast of language group.

Brain region	Jpn – Eng Accented + Unaccented Figure 2H (red)	Eng – Jpn Accented + Unaccented Figure 2H (blue)
PMvi, Broca's area, BA 6,44	−45,0,8	
	48,12,9	
PMvs/PMd BA 6	−30,0,36	
	30,0,39	
	39, −15,60	
PostCG, IPL BA1,2	−60,−18,21	
	51,−24,60	
PostCG, IPL BA3		−30,−24,48
Superior medial gyrus BA10		−9,54,0
Medial frontal gyrus/SFG BA9		−30,30,24
		−15,51,39
		18,33,33
Middle frontal gyrus BA11		−30,34,−19
Anterior cingulate gyrus		9,51,15
Middle cingulate cortex BA24,31		−12,−39,42
		12,−33,45
		12,−3,45
IPL BA40	−45,−39,39	
SPL BA7		
Insula BA13, 47	36,18,−9	−36,−18,15
MTG /STG BA21,22	−51,−48,6	
Angular gyrus BA39		−54,−66,24
MOG BA18,19	−27,−69,30	−36,−87,27
Cuneus/Precuneus		−9,−72,24
		24,−63,18
Lingual gyrus BA18		−9,−60,0
Cerebellum	−27,−69,−30	
Lobule VIIa Crus I	−39,−69,−36	
	21,−66,−36	
Cerebellum Lobule V	−6,−57,−30	
Cerebellum Lobule VI	−15,−72,−27	
Putamen	30,9,0	

Table showing clusters of activity for the main effect of language group thresholded at pFDR < 0.05 corrected with an extent threshold greater than 5 voxels. Jpn., Japanese; Eng., English; Cor., corrected for multiple comparisons; BA, Brodmann area; PMvi, Ventral inferior premotor cortex; PMvs, Ventral superior premotor cortex; PMd, Dorsal premotor cortex; PostCG, Postcentral gyrus; SFG, Superior Frontal Gyrus; IPL, Inferior parietal lobule; SPL, Superior Parietal Lobule; MTG, Middle Temporal Lobe; STG, Superior Temporal Lobe; MOG, Middle Occipital Gyrus. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity.

Table 3 | MNI Coordinates of Clusters of Activity for Contrasts of Interest.

Brain region	Accented – Unaccented /rl/ Identification (Eng – Jpn) Figure 3	Accented – Unaccented /rl/ Identification(Eng) Figure 2E	Accented/rl/ Identification (Jpn – Eng) Figure 5A	Unaccented/rl/ Identification (Jpn – Eng) Figure 5B
PMvi, Broca's area, BA 6,44		−48,10,4 (−48,12,−3)	(−45,0,9) 48,12,6 (48,12,6)	(48,3,0) 48,9,15, 60,15,3
PMvs/PMd BA 6		33,−15,48	27,0,36 (27,0,36)	30,0,42 (30,0,42) (57,0,45) (39,−15,66)
Broca's Area BA 45	−51,30,6 (−54,30,3*) 54,27,18	−42,33,6 (−48,30,9)		54,21,9
IFG BA47		−45,24,−12 (−45,24,−12)		−30,21,−4
Rolandic operculum BA43				−63,−18,21 (57,−12,12)
MFG BA8		(−51,15,42)		
MFC including Pre-SMA	0,39,33**, 0,36,42	0,32,38, 0,29,50 (3,33,45)		
SMA				(−15,−6,66)
DLPFC	54,30,30			
MTG BA21				(69,−18,−6)
IPL BA 40				−45,−39,39, −30,−48,39
SPL				(−12,−51,66)
Cerebellum Lobule V				−15,−57,−30
Cerebellum Lobule VI		27,−60,−33 (27,−60,−33)	(−15,−57,−27)	21,−66,−36 (−18,−57,−30)
Cerebellum Lobule VII	6,−81,−33	(−3,−69,−30) −9,−87,−27 18,−72,−39		−27,−69,−30
Brain Stem	0,−30,−30			(6,−45,−36)

*Table showing clusters of activity for the various contrasts thresholded at $p < 0.001$ uncorrected. Coordinates in Parentheses denote those that are significant when using phonetic identification performance as a covariate of non-interest. Jpn., Japanese; Eng., English; BA, Brodmann area; PMvi, Ventral inferior premotor cortex; PMvs, Ventral superior premotor cortex; PMd, Dorsal premotor cortex; IFG, Inferior frontal gyrus; MFG, Middle frontal gyrus; MFC, Medial frontal cortex. SMA, Supplementary motor area; DLPFC, Dorsolateral Prefrontal Cortex; MTG, Middle Temporal Gyrus; IPL, Inferior Parietal Lobule; SPL, Superior parietal lobule. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity. *Cluster was not significant when thresholded at $p < 0.001$ uncorrected but was significant at $p < 0.0015$ uncorrected. **Significant at $p < 0.05$ FWE correcting for multiple comparisons across the entire volume.*

activity in left and right PMvi/Broca's, right PMvs/PMd and left and right cerebellum lobule VI (see **Figure 7** and **Table 4**). These same brain regions were shown to have significant activation (correcting for multiple comparisons) when using phonetic identification performance as a covariate of non-interest.

DISCUSSION

The goal of this study was to determine if there are differences in the level and/or patterns of activation for various brain regions involved with the processing of accented speech when distinct phonetic categories existed within a listener's language networks (first-language), relative to when listeners do not have well established phonetic categories (second-language) (i.e., English /r/ and /l/ identification for native Jpn speakers). The conjunction analysis of all four conditions [Eng(accented), Eng(unaccented), Jpn(accented), Jpn(unaccented)] revealed that the same brain regions (STG/S, MTG, SMG, Broca's area, PMC, medial frontal cortex MFC/pre-suplementary motor area, and the cerebellum lobule VI) were active (see **Figures 2A–D,G** and **Table 1**). These results suggest that, to a large extent, it is the *level* of activity within these common regions that differs between conditions, rather than recruitment of different regions in the brain. It should be noted that, even for the Eng unaccented condition, there was common activation in speech motor regions.

Increased brain activity during the presentation of accented first-language phonetic categories relative to unaccented phonetic categories [Eng(accented – unaccented)] was located primarily in the left and right cerebellum, as well as in left PMvi/Broca's area, and right PMvs/PMd (see **Figure 2E**, **Tables 3**, **4**). These results were also found when using phonetic identification performance as a covariate of non-interest. When general stimulus and subject variables were controlled for, using the contrast of Eng(accented – unaccented) – Jpn(accented – unaccented), the brain regions with significant activation included the pre-SMA, the right cerebellum, left Broca's area BA45, and the left PMvi/Broca's area (see **Figures 3**, **4**, **Tables 3**, **4**). However, when using performance as a covariate of non-interest, only left Broca's area BA45 showed significant activity (see **Tables 3**, **4**). Broca's area BA45 is thought to provide a contextual supporting role to the mirror neuron system (Arbib, 2010). PMvi/Broca's area and the cerebellum are hypothesized to be regions that instantiate the articulatory—auditory models that are involved with both speech production and perception (Callan et al., 2004a; Tourville and Guenther, 2011; Guenther and Vladusich, 2012). The left hemisphere activity

Table 4 | ROI analysis using small volume correction for contrasts of interest.

Brain region	SVC center Coordinate (8 mm radius)	Accented – Unaccented /rl/ identification (Eng – Jpn) Figure 4		Accented – Unaccented /rl/ Identification (Eng) Figure 2E		Accented /rl/ Identification (Jpn – Eng) Figure 6		Unaccented /rl/ Identification (Jpn – Eng) Figure 7	
		pCor.	x,y,z	pCor.	x,y,z	pCor.	x,y,z	pCor.	x,y,z
PMvi, Broca's BA6,44	−51,9,21	0.030	−48,12,21	0.092	−57,6,21	n.s.	–	0.042	−54,9,27
	51,15,18	n.s.	–	n.s.	–	0.045	48,9,15	0.006	48,9,15
	CovPerf								
	−51,9,21	n.s.	–	n.s.	–	n.s.	–	n.s.	–
	51,15,18	n.s.	–	n.s.	–	0.074	48,12,12	n.s.	–
PMvs	−36,−3,57	n.s.	–	0.081	−36,0,51	n.s.	–	n.s.	–
	27,−3,51	n.s.	–	n.s.	–	0.036	27,0,45	0.006	27,0,45
	CovPerf								
	−36,−3,57	n.s.	–	0.057	−36,0,51	n.s.	–	n.s.	–
	27,−3,51	n.s.	–		–	0.063	27,0,45	0.027	21,−6,51
STG/S	−57,−39,9	n.s.	–	0.075	−57,−36,3	n.s.	–	n.s.	–
	CovPerf								
	−57,−39,9	n.s.	–	0.091	−57,−36,3	n.s.	–	n.s.	n.s.
CerebellumLobule VI	−27,−63,−39	n.s.	–	0.011	−30,−57,−36	0.042	−27,−66,−33	0.011	−27,−66,−33
	30,−66,−33	0.034	36,−72,−33	0.025	27,−60,−33	n.s.	–	0.028	24,−69,−33
	CovPerf								
	−27,−63,−39	n.s.	–	0.012	−30,−57,−36	0.039	−27,−66, −33	0.005	−21,−60,−39
	30,−66,−33	n.s.	–	0.024	27,−60,−33	n.s.	–	0.033	33,−63,−39

Table showing results of small volume correction analysis ($p < 0.05$) for multiple comparisons for selected contrasts within regions of interest using MNI coordinates specified in Callan et al. (2004a) as the seed voxels. The first set of results is for the original analysis. The second set of results, under the heading of CovPerf, is for the analysis in which phonetic identification performance is used as a covariate of non-interest. SVC, Small volume correction; ROI, Region of Interest; BA, Brodmann area; PMvi, Premotor cortex ventral inferior; PMvs, Premotor cortex ventral superior. n.s., Not significant at $p < 0.05$ corrected. pCor., p corrected for multiple comparisons within the SVC small volume corrected region of interest. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity.

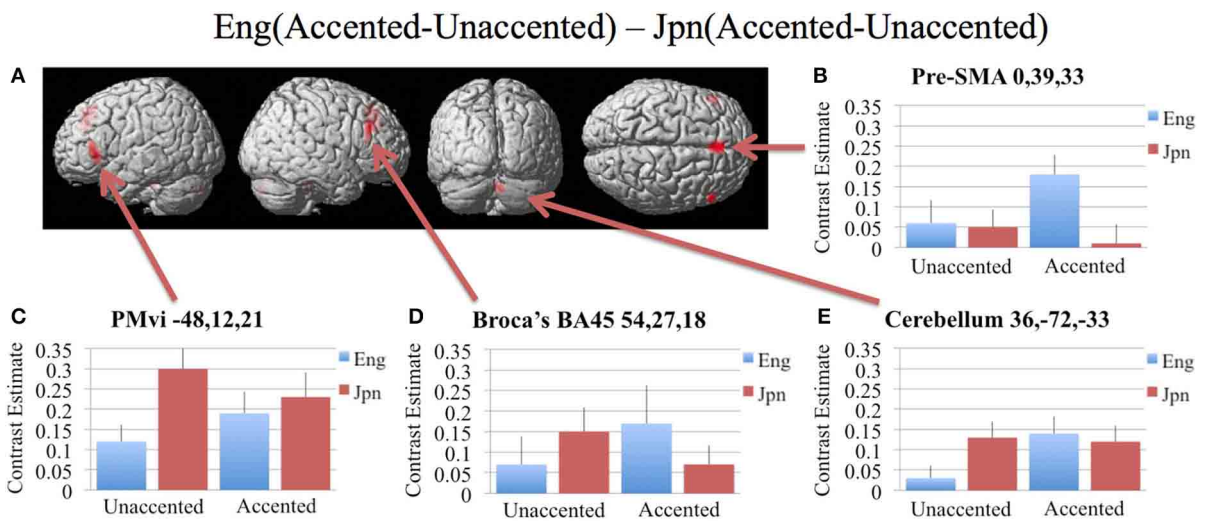


FIGURE 3 | Significant brain activity (thresholded at $p < 0.001$ uncorrected) for the interaction of language group and accent. This contrast focused on the activity involved with perception of foreign-accented productions of a first-language phonetic category. **(A)** Significant brain activity rendered on the surface of the brain for the contrast of Eng(accented-unaccented) – Jpn(accented-unaccented) showing activity in pre- supplementary motor area pre-SMA, left and right Broca's area BA45, right dorsolateral prefrontal cortex DLPFC, and left and right cerebellum. **(B–E)** shows contrast estimates and standard error of the SPM analysis relative to rest for the four conditions in selected regions: **(B)** Pre-SMA, **(C)** PMvi, **(D)** Broca's, **(E)** Cerebellum.

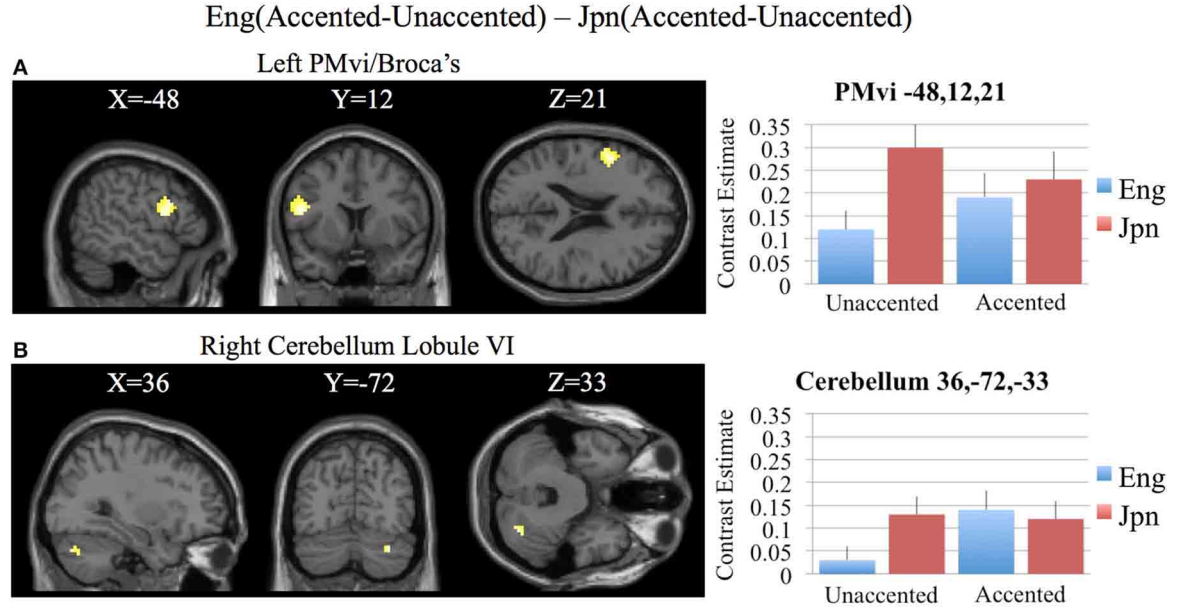


FIGURE 4 | Region of interest (ROI) analysis for the contrast of Eng(accented-unaccented) – Jpn(accented-unaccented) using small volume correction ($p < 0.05$) for multiple comparisons. (A) left PMvi/Broca's area. **(B)** Right Cerebellum Lobule VI. MNI X, Y, Z coordinates are given at the top of each brain slice. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity. The SPM contrast estimates and standard error relative to rest for all four conditions are given on the left side of each ROI rendered image.

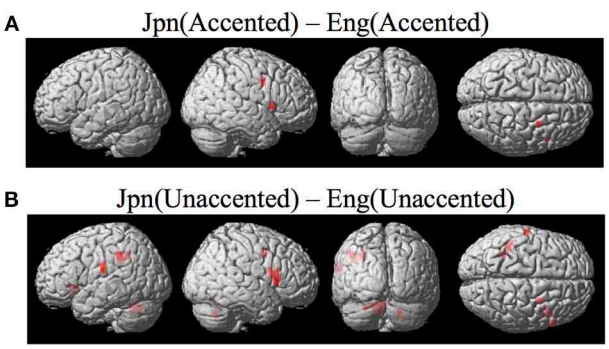


FIGURE 5 | (A) Contrast investigating specific brain regions involved with the perception of foreign-accented productions of a second-language phonetic category, Jpn(accented) – Eng(accented). Activity is present in the right ventral inferior premotor cortex including Broca's area PMvi/Broca's right ventral superior premotor cortex PMvs. **(B)** Activity for perception of foreign-accented productions of a second language phonetic category that may not be specific Jpn(unaccented) – Eng(unaccented) was found in the left and right PMvi/Broca's, the right PMvs/PMd, the right Broca's area BA45, the left inferior frontal gyrus BA47, the left postcentral gyrus, the left inferior parietal lobule, and the left and right cerebellum.

observed in Broca's area BA 45 and PMvi/Broca's area, is consistent with other studies that showed only left hemisphere activity for speech perception tasks that required phonetic processing (Demonet et al., 1992; Price et al., 1996). The presence of increased activity in speech motor regions observed in this study, and the lack of significant differential activity in the STG/S, are consistent with the hypothesis that neural processes involved with auditory—articulatory mappings are used to facilitate the perception of foreign-accented productions of one's first language. However, the absence of differential activity in auditory regions for this contrast does not indicate that auditory processes are not important for intelligibility and perceptual categorization.

The activity present in the MFC that included the pre-SMA for all conditions (see **Figure 2** and **Table 1**) is interesting given that several studies suggest that this region may be involved with value and contex-dependent selection of actions (Deiber et al., 1999; Lau et al., 2004; Rushworth et al., 2004). Activity found in the MFC/Pre-SMA in this study may represent value and context dependent selection of internal models. It is important to note that the contrast Eng (accented) vs. Jpn (accented) showed greater activity in the MFC (see **Figures 3**, **4**, **Tables 3**, **4**). This was also true when phonetic identification performance was used as a covariate of non-interest. This suggests greater use of value-dependent context for selection when internal models are well established (as is thought to be the case for /r/ and /l/ for native English speakers). This region was also displayed significant activation when the Eng vs. Jpn groups were compared (see **Figure 2H**, **Table 2**). The greater extent of activity in these regions compared to the Callan et al. (2004a) study may be explained by the larger number of speakers used for the stimuli in this study, which could have resulted in considerably more context variability.

Brain regions specific to the perception of foreign-accented productions of phonetic categories from one's second language, when controlling for task difficulty [Jpn(accented) – Eng(accented)], was localized in right PMvi/Broca's area, right

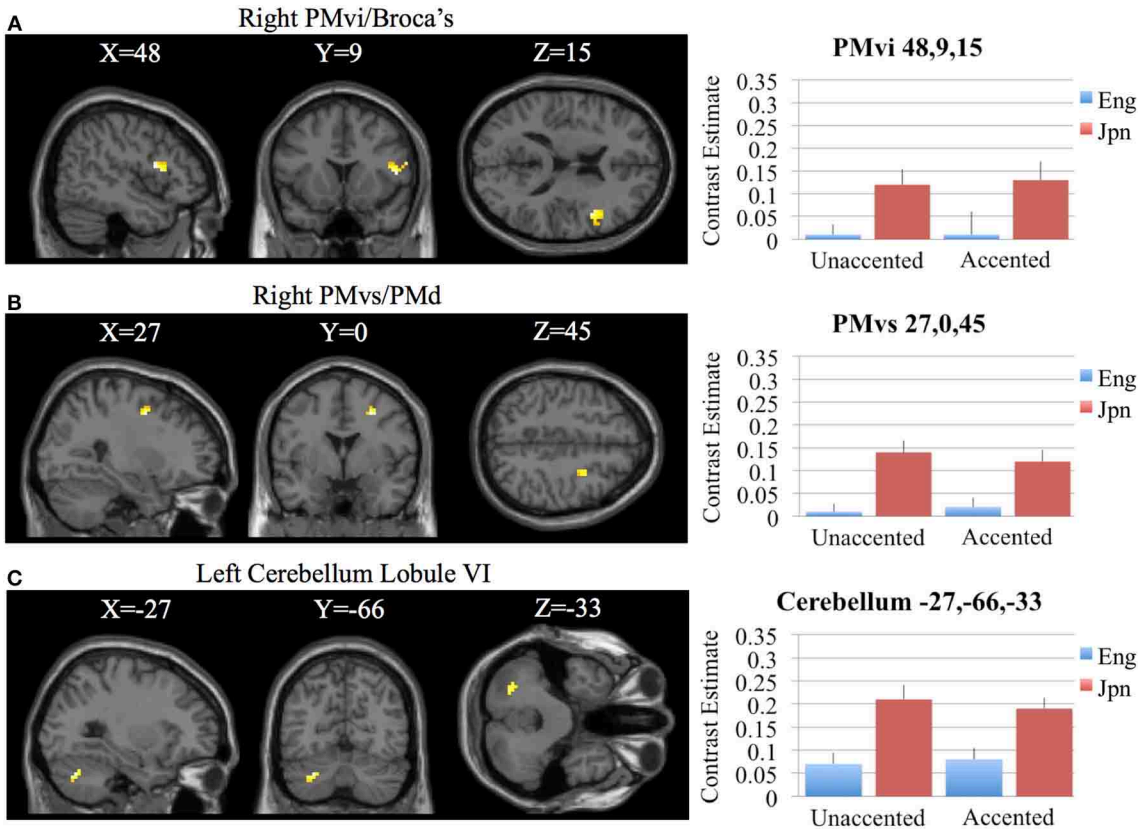


FIGURE 6 | Region of interest analysis for the contrast of Jpn(accented) – Eng(accented) using small volume correction ($p < 0.05$) for multiple comparisons. (A) Right PMvi/Broca's area. **(B)** Right PMvs/PMd. **(C)** Left Cerebellum Lobule VI. MNI X, Y, Z coordinates are given at the top of each brain slice. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity. The SPM contrast estimates and standard error relative to rest for all four conditions are given on the left side of each ROI rendered image.

PMvs/PMd, and the left cerebellum (see **Figures 5A**, **6**, and **Tables 3**, **4**). These results are also true when using phonetic identification performance as a covariate of non-interest. Task difficulty was controlled for by presenting foreign accented speech (English /rl/ phonetic contrast) that was difficult for both native English and native Japanese speakers to correctly identify. It is important to point out that behavioral performance during the fMRI experiment revealed no significant difference between native English and native Japanese speakers for the foreign accented stimuli, which suggests similar levels of task difficulty for both groups.

The contrast Jpn(unaccented) – Eng(unaccented) revealed activity in right PMvi/Broca's, right PMvs/PMd, the right and the left cerebellum (see **Figures 5B**, **7** and **Tables 3**, **4**). Activity in these regions was also present when using phonetic identification as a covariate of non-interest. The presence of activity in right PMvs/PMd for the Jpn(accented) – Eng(accented) contrast and the Jpn(unaccented) – Eng(unaccented) contrast suggests that the results found are not specific to acoustic properties inherent in accented speech. It should be noted that no significant activity was found in the STG/S, which is thought to be involved with auditory-based speech processing.

It should be acknowledged that difference in the number of men and women in the Eng and the Jpn groups may be responsible for the between-group differences reported here. However, the Eng (Accented – Unaccented) – Jpn (Accented – Unaccented) should control for such subject differences. As well, we believe that it is unlikely that gender differences between the groups contributed to our results, given that Callan et al. (2004a) did not find gender differences using a very similar task. In addition, no gender differences were found in another study that employed speech production tasks (Buckner et al., 1995).

It has been previously suggested that activity in speech motor regions (PMC and Broca's area) may not be involved with speech intelligibility, but rather reflect differences in cognitive processes related to task difficulty, such as attention and working memory (Hickok and Poeppel, 2007; Poeppel et al., 2008; Lotto et al., 2009; Scott et al., 2009). While all four of the primary contrasts investigated in this study controlled for general processes related to the phonetic categorization task, only the contrast Jpn(accented) – Eng(accented) adequately controlled for task difficulty. The other two primary contrasts of interest [Jpn(unaccented) – Eng(unaccented) and Eng(accented-unaccented) – Jpn(accented-unaccented)] did not.

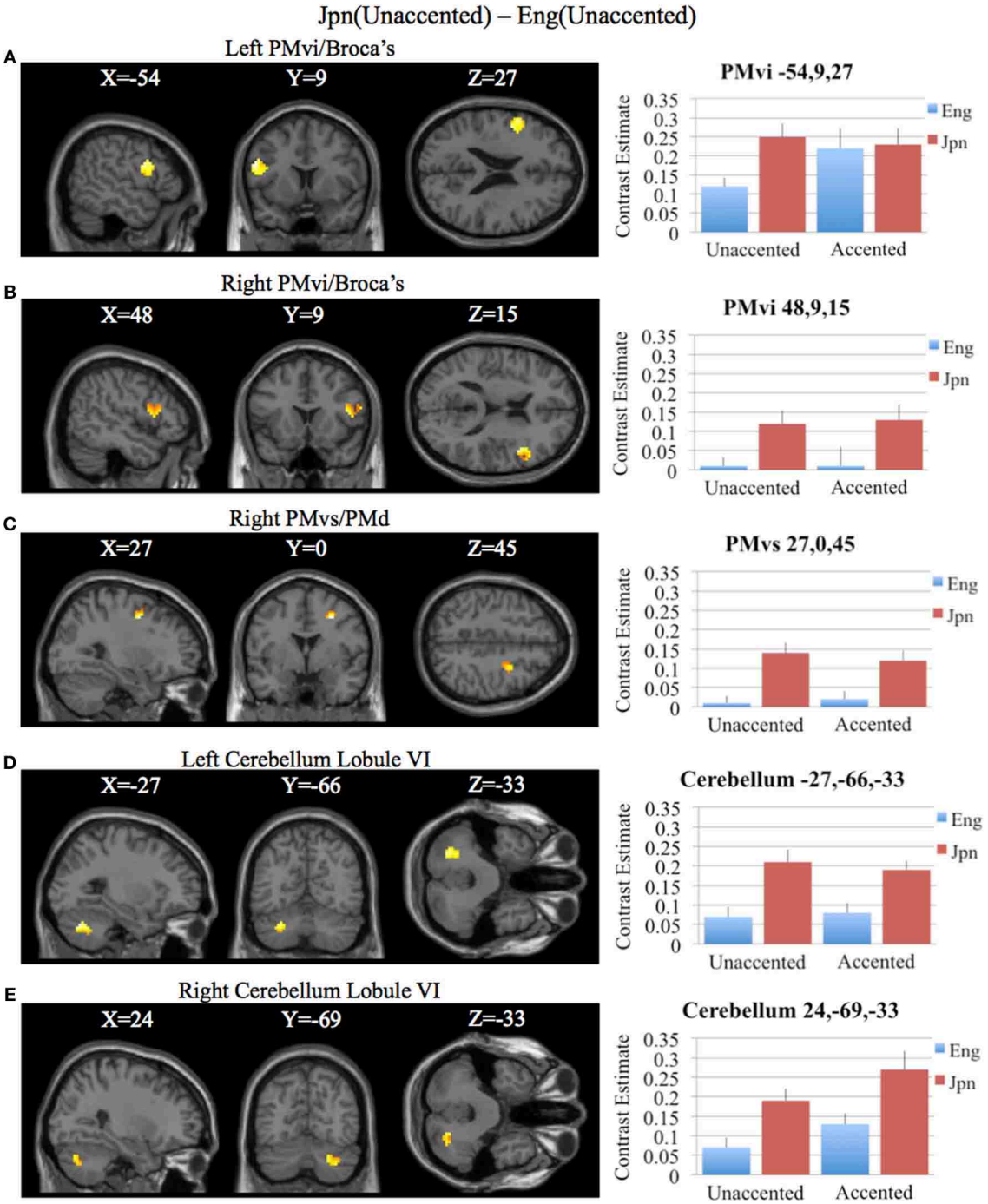


FIGURE 7 | Region of interest analysis for the contrast of Jpn(unaccented) – Eng(unaccented) using small volume correction ($p < 0.05$) for multiple comparisons. (A) Left PMvi/Broca's area. **(B)** Right PMvi/Broca's area. **(C)** Right PMvs/PMd. **(D)** Left Cerebellum Lobule VI. **(E)** Right Cerebellum Lobule VI. MNI X, Y, Z coordinates are given at the top of each brain slice. Negative × MNI coordinates denote left hemisphere and positive × values denote right hemisphere activity. The SPM contrast estimates and standard error relative to rest for all four conditions are given on the left side of each ROI rendered image.

Pertinent to the issue of controlling for extraneous brain activity related to aspects of task difficulty, the four primary contrasts in this study were analyzed using phonetic identification performance as a covariate of non-interest. The results (see **Tables 3**, **4**) showed that many of the same regions (including the PMC, Broca's area, and the cerebellum) were still found to be differentially active when performance was used as a covariate of non-interest. One drawback of using phonetic identification

performance as a covariate of non-interest to control for task difficulty is that brain activity related to the processes of enhancing speech perception is likely removed by the analysis.

Of particular interest is the finding that while the perception of foreign-accented productions of a first language is related to increased activity in left PMvi/Broca's area and the right cerebellum, brain regions involved in the perception of foreign-accented productions of a second language differentially activate right PMvs/PMd and the left cerebellum instead. While left PMvi/Broca's area is thought to be involved with articulatory and sensory aspects of phonetic processing (Guenther and Vladusich, 2012), the right premotor cortex is thought to be involved with articulatory-to-auditory mapping for feedback control (Tourville and Guenther, 2011). These results are consistent with the hypothesis that the establishment of non-native phonetic categories (when the second-language is acquired after childhood) involves greater reliance on general articulatory-to-auditory feedback control systems. These systems are thought to be instantiated in right hemisphere PMC, and generate auditory predictions based on articulatory planning (Tourville and Guenther, 2011; Guenther and Vladusich, 2012).

Selective activity in right PMC and the left cerebellum (cerebellar cortical anatomical connectivity is predominantly crossed) is consistent with the hypothesis that internal models in the non-dominant hemisphere are utilized more extensively under conditions in which there is interference between established categorical representations and new representations during processing. Some additional evidence consistent with this hypothesis comes from studies in which non-native speech training led to enhanced activity in right PMC and Broca's area (Callan et al., 2003b; Wang et al., 2003; Golestani and Zatorre, 2004) and the left cerebellum (Callan et al., 2003b). Also consistent are the results of some studies investigating second-language processing that showed greater differential activity for second-language processing than for first-language processing in right PMC and Broca's area (Dehaene et al., 1997; Pillai et al., 2003) and the left cerebellum (Pillai et al., 2004). However, there are several studies that do not show any difference in brain activity between first- and second-language processing (Klein et al., 1995; Chee et al., 1999; Illes et al., 1999). It is important to note that even though the results of this study support the hypothesis that right Broca's area and the left cerebellum are differentially involved in the processing of foreign-accented productions of a second language, left Broca's area and the right cerebellum are involved with general processing of foreign-accented phonemes for both first- and second-language listeners (see **Tables 3**, **4**). Although it is thought that the activity in the left cerebellum and right Broca's area represents articulatory-auditory internal models, it is possible that the activity represents articulatory-orosensory internal models or both articulatory-auditory and articulatory-orosensory internal models. Further experiments are needed to discern the types of internal models used under differing conditions.

The activation in left and right cerebellar lobule VI was within the region known to be involved with lip and tongue representation (Grodd et al., 2001). Given the predominantly crossed anatomical connectivity between the cerebellum and cortical areas, the finding of left PMC and right cerebellar activity that was found is consistent with the use of internal models for processing first-language phonemes. In contrast, the right PMC and left cerebellar activity that was found is consistent with the use of internal models used differentially for perception of foreign-accented productions of a second language. These results are consistent with crossed patterns of functional connectivity from the cerebellum to Broca's area that have been associated with tool use (Tamada et al., 1999). This region of the cerebellum has also been identified to be involved with speech perception and production in other studies (Ackermann et al., 2004; Callan et al., 2004a).

The finding of cerebellar activity involved in the perception of foreign-accented speech is consistent with a recent study that showed greater activity in the cerebellum after adaptation to acoustically distorted speech (Guediche et al., 2014). In contrast to our hypotheses concerning the use of forward and inverse (articulatory-auditory) internal models, Guediche et al. (2014) concluded that the cerebellum utilizes supervised learning mechanisms that rely purely on sensory prediction error signals for speech perception.

Another potential explanation of the results differentiating between processing of foreign-accented speech between first- and second-language speakers could be that there is recruitment of extra neural resources when undertaking tasks for which we are not trained. It has been shown, for example, that experienced singers, in which much of the processing is automated, show reduced activity relative to non-experienced singers (Wilson et al., 2011). It is unlikely that the results of our study can be explained by differences in task training and expertise, as the foreign-accented speech was difficult for both the English and Japanese groups, and the subjects had the same amount of training on the phonetic categorization task. As well, there was no significant difference in behavioral performance between the two groups (see **Figure 1**). However, it may be the case that very different processes are recruited when distinct phonetic categories exists (first-language perception), vs. when they do not (second-language perception). Although our results are consistent with the hypothesis that the establishment of second-language phonetic categories involves general articulatory-to-auditory feedback control systems in right hemisphere PMC—which generate auditory predictions based on articulatory planning, it cannot be ruled out that the pattern of differential activity reflects meta-cognitive processing strategies that result from the task requirement to identify phonetic categories that either are either from one's first or second-language. The processes may be more automatic for native speakers (or speakers with well-established phonetic categories) than for non-native speakers.

CONCLUSION

The results of this study suggest that perception of foreign-accented phonetic categories involves brain regions that support aspects of speech motor control. For perception of foreign-accented productions of a first language, the activation in left PMvi/Broca's area, right cerebellum lobule VI, and the pre-SMA are consistent with the hypothesis that internal models instantiating auditory-articulatory mappings of phonemes are selected to facilitate perception. Brain regions selective for perception of second-language phonetic categories include right PMvi/Broca's,

right PMvs/PMd, and the left cerebellum and are consistent with the hypothesis that articulatory-to-auditory mappings used for feedback control of speech production are used to facilitate phonetic identification. The lack of activity in the STG/S for any of the contrasts under investigation would tend to refute the hypotheses that strong engagement of bottom-up auditory processing facilitates speech perception of foreign-accented speech under these conditions. Brain regions involved with articulatory-auditory feedback for speech motor control may be a precursor for development of perceptual categories.

ACKNOWLEDGMENTS

We would like to acknowledge the help of the Department of Multilingual Learning (ATR) (Reiko Akahane-Yamada Dept. Head) as well as the fMRI technicians Yasuhiro Shimada, Ichiro Fujimoto, and Yuko Shakudo.

REFERENCES

Ackermann, H., Mathiak, K., and Ivry, R. (2004). Temporal organization of "internal Speech" as a basis for cerebellar modulation of cognitive functions. *Behav. Cogn. Neurosci. Rev.* 3, 14–22. doi: 10.1177/1534582304263251

Adank, P., Rueschemeyer, S., and Bekkering, H. (2013). The role of accent imitation in sensorimotor integration during processing of intelligible speech. *Front. Hum. Neurosci.* 7:634. doi: 10.3389/fnhum.2013.00634

Akahane-Yamada, R. (1996). "Learning non-native speech contrasts: what laboratory training studies tell us," in *Proceedings of Acoustical Society of America and Acoustical Society of Japan Third Joint Meeting* (Honolulu, HI), 953–958.

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

Arbib, M. (2010). Mirror system activity for action and language is embedded in the integraton of dorsal and ventral pathways. *Brain Lang.* 112, 12–24. doi: 10.1016/j.bandl.2009.10.001

Bradlow, A., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: long-term retention of learning in perception and production. *Percept. Psychophys.* 61, 977–985. doi: 10.3758/BF03206911

Bradlow, A., Pisoni, D., Akahane-Yamada, R., and Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ an /l/: IV. some effects of perceptual learning on speech production. *J. Acoust. Soc. Am.* 101, 2299–2310. doi: 10.1121/1.418276

Buckner, R. L., Raichle, M. E., and Petersen, S. E. (1995). Dissociation of human prefrontal cortical areas across different speech production tasks and gender groups. *J. Neurophysiol.* 74, 2163–2173.

Callan, A., Callan, D., Tajima, K., and Akahane-Yamada, R. (2006a). Neural processes involved with perception of non-native durational contrasts. *Neuroreport* 17, 1353–1357. doi: 10.1097/01.wnr.0000224774.66904.29

Callan, D., Callan, A., Gamez, M., Sato, M., and Kawato, M. (2010). Premotor cortex mediates perceptual performance. *Neuroimage* 51, 844–858. doi: 10.1016/j.neuroimage.2010.02.027

Callan, D. E., Callan, A. M., Honda, K., and Masaki, S. (2000). Single-sweep EEG analysis of neural processes underlying perception and production of vowels. *Cogn. Brain Res.* 10, 173–176. doi: 10.1016/S0926-6410(00)00025-2

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

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., and Akahane-Yamada, R. (2003b). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage* 19, 113–124. doi: 10.1016/S1053-8119(03)00020-X

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

Callan, D., Jones, J. A., Munhall, K., Callan, A., Kroos, C., and Vatikiotis-Bateson, E. (2003a). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport* 14, 2213–2218. doi: 10.1097/01.wnr.0000095492.38740.8f

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

Callan, D., Kawato, M., Parsons, L., and Turner, R. (2007). Speech and song: the role of the cerebellum. *Cerebellum* 6, 321–327. doi: 10.1080/14734220601187733

Callan, D., and Manto, M. (2013). "Cerebellar control of speech and song," in *Handbook of the Cerebellum and Cerebellar Disorders*, eds M. Manto, D. Gruol, J. Schmahmann, N. Koibuchi, and F. Rossi (New York, NY: Springer), 1191–1199.

Callan, D., Tsytsarev, V., Hanakawa, T., Callan, A., Katsuhara, M., Fukuyama, H., et al. (2006b). Song and speech: brain regions involved with perception and covert production. *Neuroimage* 31, 1327–1342. doi: 10.1016/j.neuroimage.2006.01.036

Chee, M. W., Tan, E. W., and Thiel, R. (1999). Mandarin and English single word processing studied with functional magnetic resonance imaging. *J. Neurosci.* 19, 3050–3056.

Davachi, L., Maril, A., and Wagner, A. D. (2001). When keeping in mind supports later bringing to mind: neural markers of phonologi- cal rehearsal predict subsequent remembering. *J. Cogn. Neurosci.* 13, 1059–1070. doi: 10.1162/089892901753294356

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., et al. (1997). Anatomical variability in the cortical representation of first and second language. *Neuroreport* 8, 3809–3815. doi: 10.1097/00001756-199712010-00030

Deiber, M., Honda, M., Ibanez, V., Sadato, N., and Hallett, M. (1999). Mesial motor areas in self-initiated versus externally triggered movements examined with fMRI: effect of movement type and rate. *J. Neurophysiol.* 81, 3065–3077.

Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., Wise, R. S., et al. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain* 115, 1753–1768. doi: 10.1093/brain/115.6.1753

Golestani, N., and Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Goslin, J., Duffy, H., and Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain Lang* 122, 92–102. doi: 10.1016/j.bandl.2012.04.017

Grodd, W., Hulsmann, E., Lotze, M., Wildgruber, D., and Erb, M. (2001). Sensorimotor mapping of the human cerebellum: fMRI evidence of somatotopic organization. *Hum. Brain Mapp.* 13, 55–73. doi: 10.1002/hbm.1025

Guediche, S., Holt, L., Laurent, P., Lim, S., and Fiez, J. (2014). Evidence for cerebellar contributions to adaptive plasticity in speech perception. *Cereb. Cortex*. doi: 10.1093/cercor/bht428. [Epub ahead of print].

Guenther, F., and Vladusich, T. (2012). A neural theory of speech acquisition and production. *J. Neurolinguistics* 25, 408–422. doi: 10.1016/j.jneuroling.2009.08.006

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

Iacoboni, M. (2008). The role of premotor cortex in speech perception: evidence from fMRI and rTMS. *J. Physiol. Paris* 102, 31–34. doi: 10.1016/j.jphysparis.2008.03.003

Iacoboni, M., and Wilson, S. (2006). Beyond a single area: motor control and language within a neural architecture encompassing Broca's area. *Cortex* 42, 503–506. doi: 10.1016/S0010-9452(08)70387-3

Illes, J., Francis, W. S., Desmond, J. E., Gabrieli, J. D., Glover, G. H., Poldrack, R., et al. (1999). Convergent cortical representation of semantic processing in bilinguals. *Brain Lang.* 70, 347–363. doi: 10.1006/brln.1999.2186

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Putz, B., et al. (2000). Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* 403, 192–195. doi: 10.1038/35003194

Jonides, J., Schumacher, E. H., Smith, E. E., Koeppe, R. A., Awh, E., Reu- ter-Lorenz, P. A., et al. (1998). The role of pa- rietal cortex in verbal working memory. *J. Neurosci.* 18, 5026–5034.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* 9, 718–727. doi: 10.1016/S0959-4388(99)00028-8

Klein, D., Milner, B., Zatorre, R. J., Meyer, E., and Evans, A. C. (1995). The neural substrates underlying word generation: a bilingual functional imaging study. *Proc. Nat. Acad. Sci. U.S.A.* 92, 2899–2903. doi: 10.1073/pnas.92.7.2899

Lau, H. C., Rogers, R. D., Ramnani, N., and Passingham, R. E. (2004). Willed action and attention to the selection of action. *Neuroimage* 21, 1407–1415. doi: 10.1016/j.neuroimage.2003.10.034

Lively, S., Pisoni, D., Yamada, R., Tohkura, Y., and Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *J. Acoust. Soc. Am.* 96, 2076–2087. doi: 10.1121/1.410149

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

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

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J. J., and Fujimura, O. (1975). An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. *Percept. Psychophys*. 18, 331–340. doi: 10.3758/BF03211209

Moulin-Frier, C., and Arbib, M. (2013). Recognizing speech in a novel accent: the motor teory of speech perception reframed. *Biol Cybern.* 107, 421–447. doi: 10.1007/s00422-013-0557-3

Nishitani, N., Schurmann, M., Amunts, K., and Hari, R. (2005). Broca's region: from action to language. *Physiology* 20, 60–69. doi: 10.1152/physiol.00043.2004

Pillai, J. J., Allison, J. D., Sethuraman, S., Araque, J. M., Thiruvaiyaru, D., Ison, C. B., et al. (2004). Functional MR imaging study of language-related differences in bilingual cerebellar activation. *Am. J. Neuroradiol.* 25, 523–532. doi: 10.1016/S1053-8119(03)00151-4

Pillai, J. J., Araque, J. M., Allison, J. D., Sethuraman, S., Loring, D. W., Thiruvaiyaru, D., et al. (2003). Functional MRI study of semantic and phonological language processing in bilingual subjects: preliminary findings. *Neuroimage* 19, 565–576. doi: 10.1016/S1053-8119(03)00151-4

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.* 12, 363, 1071–86. doi: 10.1098/rstb.2007.2160

Poldrack, R. (2000). Imaging brain plasticity: conceptual and methodological issues – a theoretical review. *Neuroimage* 12, 1–13. doi: 10.1006/nimg.2000.0596

Price, C. J., Wise, R. J., Warburton, E. A., Moore, C. J., Howard, D., Patterson, K., et al. (1996). Hearing and saying: the functional neuro-anatomy of auditory word processing. *Brain* 119, 919–931. doi: 10.1093/brain/119.3.919

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

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

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., and Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends Cogn. Sci.* 8, 410–417. doi: 10.1016/j.tics.2004.07.009

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

Schmahmann, J., Doyon, J., Toga, A. W., Petrides, M., and Evans, A. C. (2000). *MRI Atlas of the Human Cerebellum*. San Diego, CA: Academic Press.

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

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

Skipper, J., Goldin-Meadow, S., Nusbaum, H., and Small, S. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain Lang.* 101, 260–277. doi: 10.1016/j.bandl.2007.02.008

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111, 1872–1891. doi: 10.1121/1.1458026

Strange, W., and Jenkins, J. J. (1978). "Role of linguistic experience in the perception of speech," in *Perception and Experience*, eds R. D. Walk and H. L. Pick (New York, NY: Academic), 125–169.

Talairach, J., and Tournoux, P. (1988). *Co-planar Stereotactic Atlas of the Human Brain*. New York, NY: Thieme.

Tamada, T., Miyauchi, S., Imamizu, H., Yoshioka, T., and Kawato, M. (1999). Cerebro-cerebellar functional connectivity revealed by the laterality index in tool-use learning. *Neuroreport* 10, 325–331. doi: 10.1097/00001756-199902050-00022

Tourville, J., and Guenther, F. (2011). The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* 26, 952–981. doi: 10.1080/01690960903498424

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Dev.* 47, 466–472. doi: 10.2307/1128803

Wang, Y., Sereno, J. A., Jongman, A., and Hirsch, J. (2003). fMRI Evidence for cortical modification during learning of mandarin lexical tone. *J. Cogn. Neurosci.* 15, 1019–1027. doi: 10.1162/089892903770007407

Werker, J. F., Gilbert, J. H. V., Humphrey, K., and Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Dev*. 52, 349–355. doi: 10.2307/1129249

Werker, J. F., and Tees, R. C. (1999). Influences on infant speech processing: toward a new synthesis. *Annu. Rev. Psychol.* 50, 509–535. doi: 10.1146/annurev.psych.50.1.509

Wilson, S. J., Abbott, D., Lusher, D., Gentle, E., and Jackson, G. (2011). Finding your voice: a singing lesson from functional imaging. *Hum. Brain Mapp.* 32, 2115–2130. doi: 10.1002/hbm.21173

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

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

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

Received: 31 January 2014; accepted: 14 August 2014; published online: 03 September 2014.

*Citation: Callan D, Callan A and Jones JA (2014) Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. Front. Neurosci. **8**:275. doi: 10.3389/fnins.2014.00275*

This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

REVIEW ARTICLE
published: 03 June 2014
doi: 10.3389/fnins.2014.00132

How learning to abstract shapes neural sound representations

Anke Ley[1,2], Jean Vroomen[1] and Elia Formisano[2]*

[1] Department of Medical Psychology and Neuropsychology, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, Netherlands
[2] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Rajeev D. S. Raizada, Cornell University, USA
Andre Brechmann, Leibniz Institute for Neurobiologie, Germany

***Correspondence:**
Elia Formisano, Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, PO Box 616, 6200 MD Maastricht, Netherlands
e-mail: e.formisano@maastrichtuniversity.nl

The transformation of acoustic signals into abstract perceptual representations is the essence of the efficient and goal-directed neural processing of sounds in complex natural environments. While the human and animal auditory system is perfectly equipped to process the spectrotemporal sound features, adequate sound identification and categorization require neural sound representations that are invariant to irrelevant stimulus parameters. Crucially, what is relevant and irrelevant is not necessarily intrinsic to the physical stimulus structure but needs to be learned over time, often through integration of information from other senses. This review discusses the main principles underlying categorical sound perception with a special focus on the role of learning and neural plasticity. We examine the role of different neural structures along the auditory processing pathway in the formation of abstract sound representations with respect to hierarchical as well as dynamic and distributed processing models. Whereas most fMRI studies on categorical sound processing employed speech sounds, the emphasis of the current review lies on the contribution of empirical studies using natural or artificial sounds that enable separating acoustic and perceptual processing levels and avoid interference with existing category representations. Finally, we discuss the opportunities of modern analyses techniques such as multivariate pattern analysis (MVPA) in studying categorical sound representations. With their increased sensitivity to distributed activation changes—even in absence of changes in overall signal level—these analyses techniques provide a promising tool to reveal the neural underpinnings of perceptually invariant sound representations.

Keywords: auditory perception, perceptual categorization, learning, plasticity, MVPA

SOUND PERCEPTION—MORE THAN TIME-FREQUENCY ANALYSIS

Despite major advances in the past years to unravel the functional organization principles of the auditory system, the neural processes underlying sound perception are still far from being understood. Complementary research in animals and humans has revealed the properties of responses of neurons and neuronal populations along the auditory pathway from the cochlear nucleus to the cortex. Current knowledge on the neural representation of the spectrotemporal features of the incoming sound is such that the sound spectrogram can be accurately reconstructed from neuronal population responses (Pasley et al., 2012). Yet, the precise neural representation of the acoustic sound features alone cannot explain sound perception fully. In fact, how a sound is perceived may be invariant to changes of its acoustic properties. Unless the context in which a sound is repeated is absolutely identical to the first encounter—which is rather unlikely under natural circumstances—recognizing a sound is not trivial, given that the acoustic properties of the two repetitions may not entirely match. Obviously, this poses an extreme challenge to the auditory system. To maintain processing efficiency, acoustically different sounds must be mapped onto the same perceptual representation. Thus, an essential part of sound processing is the reduction or perceptual categorization of the vast diversity of spectrotemporal events into meaningful (i.e., behaviorally relevant) units. However, despite the ease with which humans generally accomplish this task, the detection of relevant and invariant information in the complexity of the sensory input is not straightforward. This is also reflected in the performance of artificial voice and speech recognition systems for human-computer interaction, that is far below that of humans, which is mainly due to the difficulty of dealing with the naturally occurring variability in speech signals (Benzeguiba et al., 2007). In humans, the need for perceptual abstraction in everyday functioning manifests itself in pathological conditions such as the autism spectrum disorder (ASD). Next to their susceptibility to more general cognitive deficits in abstract reasoning and concept formation (Minshew et al., 2002), individuals with ASD tend to show enhanced processing of detailed acoustic information while processing of more complex and socially relevant sounds such as speech may be diminished (reviewed in Ouimet et al., 2012).

Speech sounds have been widely investigated in the context of sensory-perceptual transformation as they represent a prominent example of perceptual sound categories that comprise a large number of acoustically different sounds. Interestingly, there is not a clear boundary between two phoneme categories such as

/b/ and /d/: the underlying acoustic features vary smoothly from one category to the next (**Figure 1A**). Remarkably though, if people are asked to identify individual sounds randomly taken from this spectrotemporal continuum as either /b/ or /d/ their percept does not vary gradually as suggested by the sensory input. Instead, the sounds from the first portion of the continuum are robustly identified as /b/, while the sounds from the second part are perceived as /d/ with an abrupt perceptual switch in between (**Figure 1B**). Performance on discrimination tests further suggests that people are fairly insensitive to the underlying variation of the stimuli within one phoneme category, mapping various physically different stimuli onto the same perceptual object (Liberman et al., 1957). At the category boundary, however, the same extent of physical difference is perceived as a change in stimulus identity. This difference in perceptual discrimination also affects speech production, which strongly relies on online monitoring of auditory feedback. Typically, a self-produced error in the articulation of a speech sound is instantaneously corrected for if, e.g., the output vowel differs from the intended vowel category. An acoustic deviation of the same magnitude and direction may however be tolerated if the produced sound and the intended sound fall within the same perceptual category (Niziolek and Guenther, 2013). This suggests that the within-category differences in the physical domain are perceptually compressed to create a robust representation of the phoneme category while between-category differences are perceptually enhanced to rapidly detect the relevant change of phoneme identity. This phenomenon is termed "Categorical Perception" (CP, Harnad, 1987) and has been demonstrated for stimuli from various natural domains apart from speech, such as music (Burns and Ward, 1978), color (Bornstein et al., 1976; Franklin and Davies, 2004) and facial expressions of emotion (Etcoff and Magee, 1992), not only for humans but also for monkeys (Freedman et al., 2001, 2003), chinchillas (Kuhl and Miller, 1975), songbirds (Prather et al., 2009), and even crickets (Wyttenbach et al., 1996). Thus, the formation of discrete perceptual categories from a continuous physical signal seems to be a universal reduction mechanism to deal with the complexity of natural environments.

Several recent reviews have discussed the neural representation of sound categories in auditory cortex (AC) and the role of learning-induced plasticity (e.g., Nourski and Brugge, 2011; Spierer et al., 2011). The emphasis of the current review lies on recent empirical studies using natural or artificial sounds and experimental paradigms that enable separating acoustic and perceptual processing levels and avoid interference with existing category representations (such as for speech). Additionally, we discuss the opportunities of modern analyses techniques such as multivariate pattern analysis (MVPA) in studying categorical sound representations.

THE ROLE OF EXPERIENCE IN THE FORMATION OF PERCEPTUAL CATEGORIES

While CP has been demonstrated many times for a large variety of stimuli, the mechanisms underlying this phenomenon remain debated. Even for speech, which has most widely been investigated, the relative contribution of innate processes and learning in the formation of phoneme categories is not completely resolved. Despite the striking consistency of perceptual phoneme boundaries across different listeners, behavioral evidence suggests that those boundaries are malleable depending on the context in which the sounds are perceived (Benders et al., 2010). Additionally, cross-cultural studies have shown that language learning influences the discriminability of speech sounds, such that phonemes in one particular language are only perceived categorically by speakers of that language and continuously otherwise (Kuhl et al., 1992). Similarly, lifelong (e.g., musical training) as well as short-term experience both affect behavioral processing—and neural encoding (see below)—of relevant speech cues, such as pitch, timber and timing (Kraus et al., 2009). In support of the claim that speech CP can be acquired through training stand experimental learning studies that successfully induced discontinuous perception of a non-native phoneme continuum through elaborate category training (Myers and Swan, 2012). Nevertheless, even after extensive training, non-native phoneme contrasts tend to remain less robust than speech categories in the native language. Apart from the age of acquisition, the complexity of the learning environment and in particular the offered stimulus variability during category learning seems to affect the ability to discriminate novel phonetic contrasts (Logan et al., 1991). A prevalent theory for the formation of speech categories in particular is the motor theory of speech perception (Liberman and

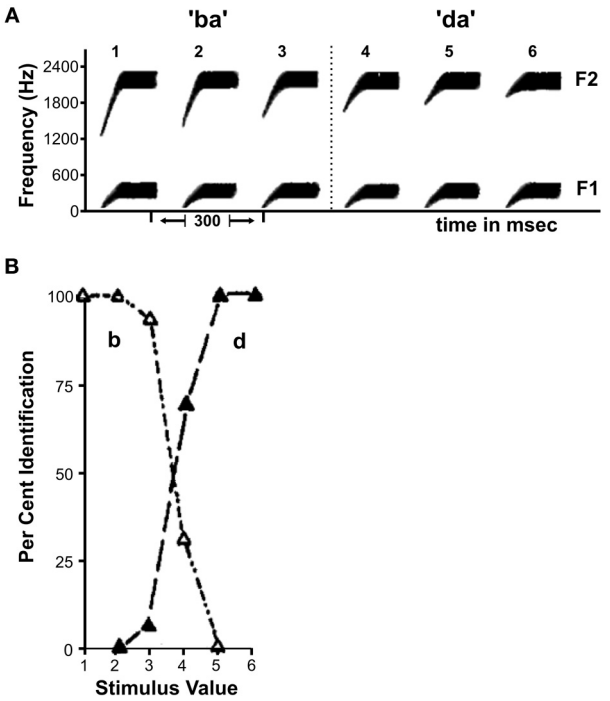


FIGURE 1 | Illustration of the sensory-perceptual transformation of speech sounds. (A) Schematic representation of spectral patterns for the continuum between the phonemes /b/ and /d/. F1 and F2 reflect the first and second formant (i.e., amplitude peaks in the frequency spectrum). **(B)** Phoneme identification curves corresponding to the continuum in A. Curves are characterized by relatively stable percepts within a phoneme category and sharp transitions in between. Figure adapted from Liberman et al. (1957).

Mattingly, 1985). This theory claims that speech sounds are categorized based on the distinct motor commands for the vocal tract used for pronunciation. Further fueled by the discovery of mirror neurons, the theory still has its proponents (for review see Galantucci et al., 2006), however, today, it is disputed in its strict form in which speech processing is considered special, as the recruitment of the motor system for sound identification has been demonstrated for various forms of non-speech action-related sounds (Kohler et al., 2002). Furthermore, accumulating evidence indicates that CP can be induced by learning for a variety of non-speech stimulus material (e.g., simple noise sounds, Guenther et al., 1999 and inharmonic tone complexes, Goudbeek et al., 2009). The use of artificially constructed categories for studying CP has the advantage that the physical distance between neighboring stimuli can be controlled such that the similarity ratings of within- or between-category stimuli can be attributed to true perceptual effects, rather than the metrics of the stimulus dimensions. Nevertheless, one should bear in mind that the long-term exposure to statistical regularities of the acoustics of natural sounds might exert a lasting influence on the formation of new sound categories. In support of this claim, Scharinger et al. (2013b) revealed a strong preference for negatively correlated spectral dimensions typical for speech and other natural categories when participants learned to categorize novel auditory stimuli. In line with this behavioral documentation in humans, a recent study in rodent pups demonstrated the proneness of auditory receptive fields to the systematics of the acoustic environment shaping the tuning curves of cortical neurons. Most importantly, these neuronal changes were shown to parallel an increase in perceptual discrimination of the employed sounds, which points to a link between (early) neuronal plasticity and perceptual discrimination ability (Köver et al., 2013). In sum, these experiments demonstrated that the perceptual abilities could be modified by learning and experience, while the role of pre-existing (i.e., innate) neural structures and their early adaptation in critical phases of maturation might play a vital role.

NEURAL REPRESENTATIONS OF PERCEPTUAL SOUND CATEGORIES

Behavioral studies have been complemented with research on the neural implementation of perceptual sound categories. Forming new sound categories or assigning a new stimulus to an existing category requires the integration of bottom-up stimulus driven information with knowledge from prior experience and memory as well as linking this information to the appropriate response in case of an active categorization task. Different research lines have highlighted the contribution of neural structures along the auditory pathway and in the cortex to this complex and dynamic process.

Functional neuroimaging studies employing natural sound categories such as voices, speech, and music have located object-specific processing units in higher level auditory areas in the superior temporal lobe (Belin et al., 2000; Leaver and Rauschecker, 2010). Particularly, native phoneme categories were shown to recruit the left superior temporal sulcus (STS) (Liebenthal et al., 2005) and the activation level of this region seems to correlate with the degree of categorical processing (Desai et al., 2008). While categorical processes in the STS were documented by further studies, the generalization to other sound categories beyond speech remains controversial, given that the employed stimuli were either speech sounds or artificial sounds with speech-like characteristics (Leech et al., 2009; Liebenthal et al., 2010). Even if speech sounds are natural examples of the discrepancy between sensory and perceptual space, the results derived from these studies may not generalize to other categories, as humans are processing experts for speech (similar to faces) even prior to linguistic experience (Eimas et al., 1987). In addition, regions in the temporal lobe were shown to retain the sensitivity to acoustic variability within sound categories, while highly abstract phoneme representations (i.e., invariant to changes within one phonetic category) appear to depend on decision-related processes in the frontal lobe (Myers et al., 2009). These results are highly compatible with those from cell recordings in rhesus monkey (Tsunada et al., 2011). Based on the analysis of single-cell responses to human speech categories, the authors suggest that "a hierarchical relationship exists between the superior temporal gyrus (STG) and the ventral PFC whereby STG provides the 'sensory evidence' to form the decision and ventral PFC activity encodes the output of the decision process." Analog to the two-stage hierarchical processing model in the visual domain (Freedman et al., 2003; Jiang et al., 2007; Li et al., 2009), the set of findings reviewed above suggests that processing areas in the temporal lobe only constitute a preparatory stage for categorization. Specifically, the model proposes that the tuning of neuronal populations in lower-level sensory areas is sharpened according to the category-relevant stimulus features, forming a task-independent reduction of the sensory input (but see below for a different view on the role of early auditory areas). In case of an active categorization task, this information is projected to higher-order cortical areas in the frontal lobe. The predominant recruitment of the prefrontal cortex (PFC) during early phases of category learning (Little and Thulborn, 2005) and in the context of an active categorization task (Boettiger and D'Esposito, 2005; Husain et al., 2006; Li et al., 2009) support the concept that it plays a major role in rule learning and attention-related processes modulating lower-level sound processing rather than being the site of categorical sound representations *per se*.

Categorical processing does however not exclusively proceed along the auditory "what" stream. To study the neural basis of CP, Raizada and Poldrack (2007) measured fMRI while subjects listened to pairs of stimuli taken from a phonetic /ba/-/da/ continuum. Responses in the supramarginal gyrus were significantly larger for pairs that included stimuli belonging to different phonetic categories (i.e., crossing the category boundary) than for pairs with stimuli from a single category. The authors interpreted these results as evidence for "neural amplification" of relevant stimulus difference and thus for categorical processing in the supramarginal gyrus. Similar analyses showed comparatively little amplification of changes that crossed category boundaries in low-level auditory cortical areas (Raizada and Poldrack, 2007). Novel findings revived the motor theory of categorical processing: Chevillet et al. (2013) provide evidence that the role of the premotor cortex (PMC) is not limited to motor-related processes during active categorization, but that the phoneme-category tuning of

premotor regions may essentially facilitate also more automatic speech processes via dorsal projections originating from pSTS. While this automatic motor route is probably limited to processing of speech and other action-related sound categories, the diversity of the categorical processing networks documented in the above cited studies demonstrates that there is not a single answer to where and how sound categories are represented. The role that early auditory cortical fields play in the perceptual abstraction from the acoustic input remains a relevant topic of current research. A recent study from Nelken's group indicated that neurons in the cat primary auditory area convey more information about abstract auditory entities than about the spectro-temporal sound structure (Chechik and Nelken, 2012). These results are in line with the proposal that neuronal populations in primary AC encode perceptual abstractions of sounds (or *auditory objects*, Griffiths and Warren, 2004) rather than their physical make up (Nelken, 2004). Furthermore, research from Scheich's group has suggested that sound representations in primary AC are largely context- and task- dependent and reflect memory-related and semantic aspects of actively listening to sounds (Scheich et al., 2007). This suggestion is also supported by the observation of semantic/categorical effects within early (∼70 ms) post-stimulus time windows in human auditory evoked potentials (Murray et al., 2006).

Finding empirical evidence for abstract categorical representations in low-level auditory cortex in humans, however, remains challenging as it requires experimental paradigms and analysis methods that allow disentangling the perceptual processes from the strong dependence of these auditory neurons on the physical sound attributes. Here, carefully controlled stimulation paradigms in combination with fMRI pattern decoding (see below) could shed light on the matter. For example, Staeren et al. (2009) were able to dissociate perceptual from stimulus-driven processes by controlling the physical overlap of stimuli within and between natural sound categories. They revealed categorical sound representations in spatially distributed and even overlapping activation patterns in early areas of human AC. Similarly, studies employing fMRI-decoding to investigate the auditory cortical processing of speech/voice categories have put forward a "constructive" role of early auditory cortical networks in the formation of perceptual sound representations (Formisano et al., 2008; Kilian-Hütten et al., 2011a; Bonte et al., 2014).

Crucially, studying context-dependence and plasticity of sound representations in early auditory areas may help unraveling their nature. For example, Dehaene-Lambertz et al. (2005) demonstrated that even early low-level sound processing is susceptible to top-down directed cognitive influences. In a combination of fMRI and electrophysiological measures, they showed that identical acoustic stimuli were processed in a different fashion, depending on the "perceptual mode" (i.e., whether participants perceived the sounds as speech or artificial whistles).

This literature review illustrates that in order to understand the neural mechanisms underlying the formation of perceptual categories, it is necessary to (1) carefully separate perceptual from acoustical sound representations, (2) distinguish between lower-level perceptual representations and higher-order or feedback-guided decision- and task-related processes and also (3) avoid interference with existing processing networks for familiar and overlearned sound categories.

LEARNING AND PLASTICITY

Most knowledge about categorical processing in the brain is derived from experiments employing speech or other natural (e.g., music) sound categories. While providing important insights about the neural representations of familiar sound categories, these studies lack the potential to investigate the mechanisms underlying the transformation from acoustic to more abstract perceptual representations. Sound processing must however remain highly plastic beyond sensitive periods early in ontogenesis to allow efficient processing adapted to the changing requirements of the acoustic environment.

Studying these rapid experience-related neural reorganizations requires controlled learning paradigms of new sound categories. With novel, artificial sounds, the acoustic properties can be controlled, such that physical and perceptual representations can be decoupled and interference with existing representations of familiar sound categories can be avoided (but see Scharinger et al., 2013b). A comparison of pre- and post-learning neural responses provides information about the amenability of sound representations along different levels of the auditory processing hierarchy to learning-induced plasticity. Extensive research by Fritz and colleagues has provided convincing evidence for learning-induced plasticity of cortical receptive fields. In ferrets that were trained on a target (tone) detection task, a large proportion of cells in primary AC showed significant changes in spectro-temporal receptive field (STRF) shape during the detection task, as compared with the passive pre-behavioral STRF. Relevant to the focus of this review, in two-thirds of these cells the changes persisted in the post-behavior passive state (Fritz et al., 2003, see also Shamma and Fritz, 2014). Additionally, recent results from animal models and human studies have revealed evidence for similar cellular and behavioral mechanisms for learning and memory in the auditory brainstem (e.g., Tzounopoulos and Kraus, 2009).

Learning studies further provide the opportunity to look into the interaction of lower-level sensory and higher-level association cortex during task- and decision-related processes (De Souza et al., 2013). In contrast to juvenile plasticity, which is mainly driven by bottom-up input, adult learning is supposedly largely dependent on top-down control (Kral, 2013). Thus, categorical processing after short-term plasticity induced by temporary changes of environmental demands might differ from the processes formed by early-onset and long-term adaptation to speech stimuli. Even though there is evidence that with increasing proficiency in category discrimination, neural processing of newly learned speech sounds starts to parallel that of native speech (Golestani and Zatorre, 2004), a discrepancy between ventral and dorsal processing networks for highly familiar native sound categories and non-native or artificial sound categories respectively has been suggested by recent work (Callan et al., 2004; Liebenthal et al., 2010, 2013). This difference potentially limits the generalization to native speech of findings derived from studies employing artificial sound categories.

Several studies have examined the changes in the neural sound representations underlying the perceptual transformations

induced by category learning. A seminal study with gerbils demonstrated that learning to categorize artificial sounds in the form of frequency sweeps resulted in a transition from a physical (i.e., onset frequency) to a categorical (i.e., up vs. down) sound representation already in the primary AC (Ohl et al., 2001). In contrast to the traditional understanding of primary AC as a feature detector, this finding implicates that sound representations at the first cortical analysis stage are more abstract and prone to plastic reorganization imposed by changes in environmental demands. In fact, sound stimuli have passed through several levels of basic feature analyses before they ascend to the superior temporal cortex (Nelken, 2004). Thus, as discussed above, sound representations in primary AC are unlikely to be faithful copies of the physical characteristics. Even though the involvement of AC in categorization of artificial sounds has also been demonstrated in humans (Guenther et al., 2004), conventional subtraction paradigms typically employed in fMRI studies lack sufficient sensitivity to demarcate distinct categorical representations. Due to the large physical variability within categories and the similarity of sounds straddling the category boundary, between-category contrasts often do not reveal significant results (Klein and Zatorre, 2011). Furthermore, the effects of category learning on sound processing as demonstrated in animals were based on changes in the spatiotemporal activation pattern without apparent changes in response strength (Ohl et al., 2001; Engineer et al., 2014). Using *in vivo* two-photon calcium imaging in mice, Bathellier et al. (2012) have convincingly shown that categorical sound representations—which can be selected for behavioral or perceptual decisions—may emerge as a consequence of non-linear dynamics in local networks in the auditory cortex (Bathellier et al., 2012, see also Tsunada et al., 2012 and a recent review by Mizrahi et al., 2014).

In human neuroimaging, these neuronal effects that do not manifest as changes in overall response levels may remain inscrutable to univariate contrast analyses. Also, fMRI designs based on adaptation, or more generally, on measuring responses to stimulus pairs/sequences (e.g., as in Raizada and Poldrack, 2007) do not allow excluding generic effects related to the processing of sound sequences or potential hemodynamic confounds, as the reflection of neuronal adaptation/suppression effects in the fMRI signals is complex (Boynton and Finney, 2003; Verhoef et al., 2008).

Modern analyses techniques with increased sensitivity to spatially distributed activation changes in absence of changes in overall signal level provide a promising tool to decode perceptually invariant sound representations in humans (Formisano et al., 2008; Kilian-Hütten et al., 2011a) and detect the neural effects of learning (**Figure 2**). Multivariate pattern analysis (MVPA) employs established classification techniques from machine learning to discriminate between different cognitive states that are represented in the combined activity of multiple locally distributed voxels, even when their average activity does not differ between conditions (see Haynes and Rees, 2006; Norman et al., 2006; Haxby, 2012 for tutorial reviews). Recently, Ley et al. (2012) demonstrated the potential of this method to trace rapid transformations of neural sound representations, which are entirely based on changes in the way the sounds are perceived induced by a few days of category learning (**Figure 3**). In their study, participants were trained to categorize complex artificial ripple sounds, differing along several acoustic dimensions into two distinct groups. BOLD activity was measured before and after training during passive exposure to an acoustic continuum spanned between the trained categories. This design ensured that the acoustic stimulus dimensions were uninformative of the trained sound categorization such that any change in the activation pattern could be attributed to a warping of the perceptual space rather than physical distance. After successful learning, locally distributed response patterns in Heschl's gyrus (HG) and its adjacency became selective for the trained category discrimination (pitch) while the same sounds elicited indistinguishable responses before. In line with recent findings in rat primary AC (Engineer et al., 2013), the similarity of the cortical activation patterns reflected the sigmoid categorical structure and correlated with perceptual rather than physical sound similarity. Thus, complementary research in animals and humans indicate that perceptual sound categories are represented in the activation patterns of distributed neuronal populations in early auditory regions, further supporting the role of the early AC in abstract and experience-driven sound processing rather than acoustic feature mapping (Nelken, 2004). It is noteworthy that these abstract categorical representations were detectable despite passive listening conditions. This is an important detail, as it demonstrates that categorical representations are (at least partially) independent of higher-order decision or motor-related processes. Furthermore, it suggests that some preparatory (i.e., multipurpose) abstraction of the physical input happens at the level of the early auditory cortex.

The mechanisms of neuroplasticity underlying category learning and the origin of the categorical organization of sound representations in the auditory cortex are still quite poorly understood and deserve further investigation. Hypotheses are primarily derived from perceptual learning studies in animals. These studies show that extensive discrimination training may elicit reorganization of the auditory cortical maps, selectively increasing the representation of the behaviorally relevant sound features (Recanzone et al., 1993; Polley et al., 2006). This suggests that environmental and behavioral demands lead to changes of the auditory tuning properties of neurons such that more neurons are tuned to the relevant features to achieve higher sensitivity in the relevant dimension. This reorganization is mediated by synaptic plasticity, i.e., the strengthening of neuronal connections following rules of Hebbian learning (Hebb, 1949; for recent review, see Caporale and Dan, 2008). Passive learning studies suggest that attention is not necessary for sensory plasticity to occur (Watanabe et al., 2001; Seitz and Watanabe, 2003). However, in contrast to the mostly unequivocal sound structure used for perceptual learning experiments, learning to categorize a large number of sounds differing along multiple dimensions requires either sound distributions indicative of the category structure (Goudbeek et al., 2009) or a task including response feedback in order to extract the relevant and category discriminative sound feature. This selective enhancement of features requires some top-down gating mechanism. Attention can act as such a filter, increasing feature saliency (Lakatos et al., 2013) by selectively modulating the tuning properties of neurons in the auditory cortex, eventually

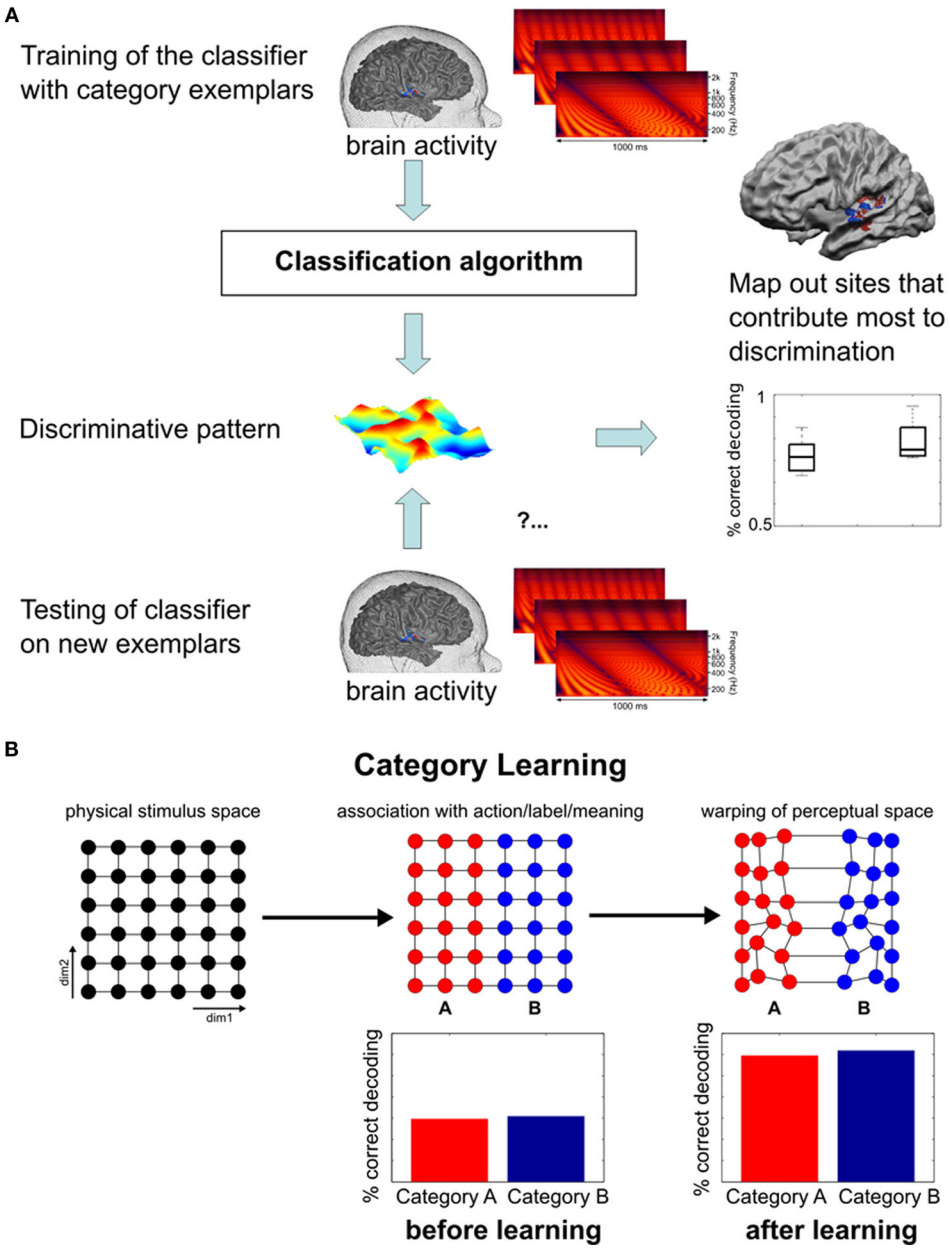


FIGURE 2 | Functional MRI pattern decoding and rationale for its application in the neuroimaging of learning. (A) General logic of fMRI pattern decoding (Figure adapted from Formisano et al., 2008). Trials (and corresponding multivariate responses) are split into a training set and a testing set. On the training set of data, response patterns that maximally discriminate the stimulus categories are estimated; the testing set of data is then used to measure the correctness of discrimination of new, unlabeled trials. For statistical assessment, the same analysis is repeated for different splits of learning and test sets. **(B)** Schematic representation of the perceptual (and possibly neural) transformation from a continuum to a discrete categorical representation. The first plot depicts an artificial two-dimensional stimulus space without physical indications of a category boundary (exemplars are equally spaced along both dimensions). During learning, stimuli are separated according to the relevant dimension, irrespective of the variability in the second dimension. Lasting differential responses for the left and right half of the continuum eventually lead to a warping of the perceptual space in which within-category differences are reduced and between-category differences enlarged. Graphics inspired by Kuhl (2000). Thus, in cortical regions where (sound) categories are represented, higher fMRI-based decoding accuracy of responses to stimuli from the two categories is expected *after learning.*

leading to a competitive advantage of behaviorally relevant information (Bonte et al., 2009, 2014; Ahveninen et al., 2011). As a consequence, more neural resources would be allocated to the behaviorally relevant information at the expense of information that is irrelevant for the decision. The adaptive allocation of neural resources to diagnostic information after category learning is supported by evidence from monkey electrophysiology (Sigala and Logothetis, 2002; De Baene et al., 2008) and human imaging,

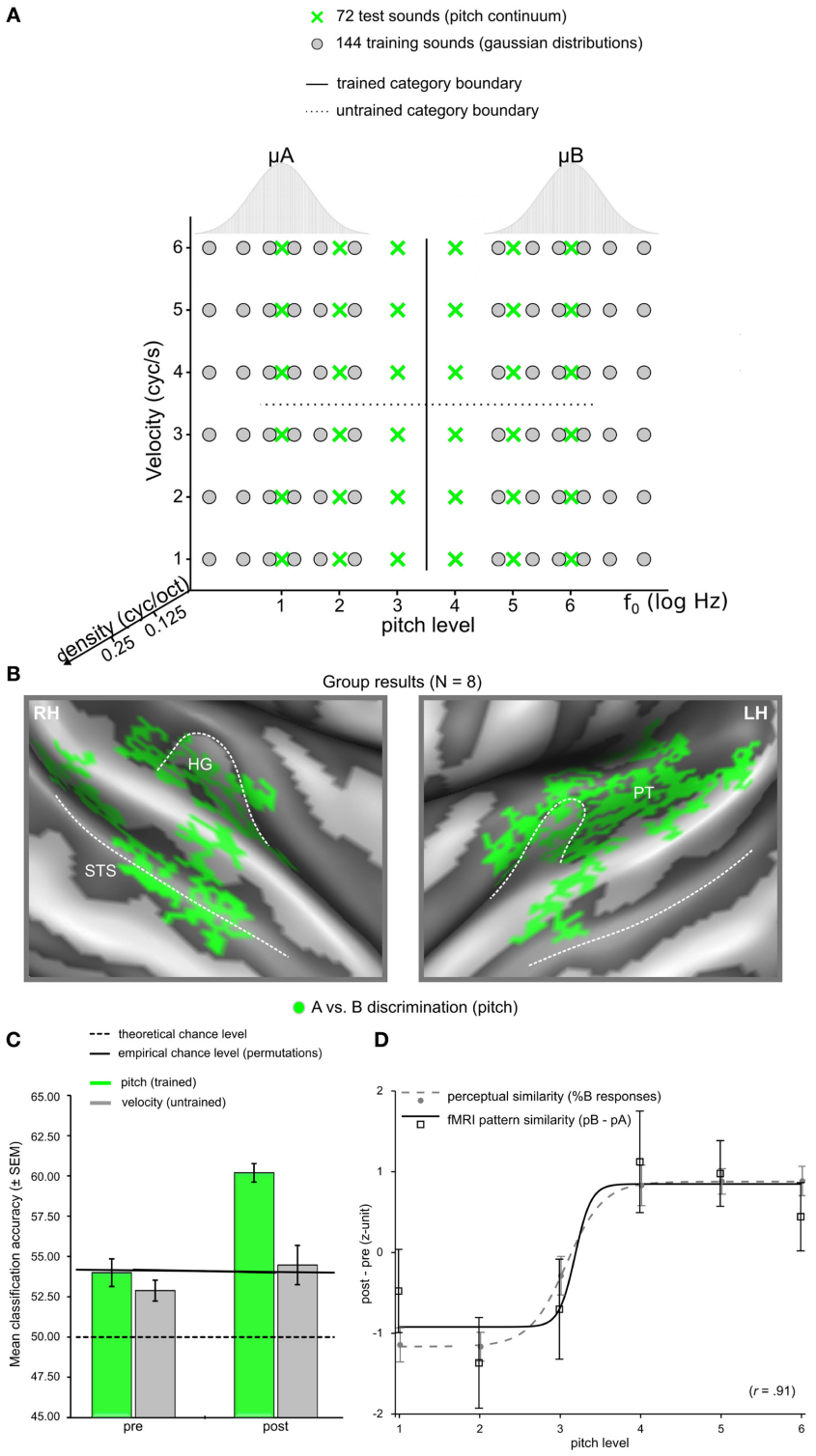


FIGURE 3 | Representation of the study by Ley et al. (2012). (A) Multidimensional stimulus space spanning the two categories A and B. **(B)** Group discrimination maps based on the post-learning fMRI data for the trained stimulus division (i.e., "low pitch" vs. "high pitch"), displayed on an average reconstructed cortical surface after cortex-based realignment. **(C)** Average classification accuracies based on fMRI data prior to category training and after successful category learning for the two types of stimulus space divisions (trained vs. untrained) and the respective trial labeling. **(D)** Changes in pattern similarity and behavioral identification curves. After category learning, neural response patterns for sounds with higher pitch (pitch levels 4, 5, 6) correlated with the prototypical response pattern for class B more strongly than class A, independent of other acoustic features. The profile of these correlations on the pitch continuum closely reflected the sigmoid shape of the behavioral category identification function.

showing decreased activation for prototypical exemplars of a category relative to exemplars near the category boundary (Guenther et al., 2004). This idea of categorical sound representations being sparse or parsimonious is also compatible with fMRI observations by Brechmann and Scheich (2005), showing an inverse correlation of auditory cortex activation and performance in an auditory categorization task. The recent discovery of a positive correlation between gray matter probability in parietal cortex and the optimal utilization of acoustic features in a categorization task (Scharinger et al., 2013a) provides further evidence for the crucial role of attentional processes in feature selection necessary for category learning. Reducing the representation of a large number of sounds too few relevant features presents an enormous processing advantage. It facilitates the read-out of the categorical pattern due to the pruned data structure and limits the neural resources by avoiding redundancies in the representation according to the concept of sparse coding (Olshausen and Field, 2004).

To date, there are several models for describing the neural circuitry between sensory and higher-order attentional processes mediating learning-induced plasticity. Predictive coding models propose that the dynamic interaction between bottom-up sensory information and top-down modulation by prior experience shapes the perceptual sound representation (Friston, 2005). This implies that categorical perception would arise from the continuous updating of the internal representation during learning to incorporate all variability present within a category, with the objective of reducing the prediction error (i.e., the difference between sensory input and internal representation). Consequently, lasting interaction between forward driven processing and backward modulation could induce synaptic plasticity and result in an internal representation that correctly matches the categorical structure and therefore optimally guides correct behavior also beyond the scope of the training period. The implementation of these Bayesian processing models rests on fairly hierarchical structures consisting of forward, backward and lateral connections entering different cortical layers (Felleman and Van Essen, 1991; Hackett, 2011). According to the Reverse Hierarchy Theory (Ahissar and Hochstein, 2004), category learning would be initiated by high-level processes involved in rule-learning, controlling via top-down modulation selective plasticity at lower-level sensory areas sharpening the responses according to the learning rule (Sussman et al., 2002; Myers and Swan, 2012). In accordance with this view, attentional modulation involving a fronto-parietal network of brain areas appears most prominent during early phases of learning, progressively decreasing with expertise (Little and Thulborn, 2005; De Souza et al., 2013). Despite recent evidence for early sensory-perceptual abstraction mechanisms in human auditory cortex (Murray et al., 2006; Bidelman et al., 2013), it is crucial to note that the reciprocal information exchange between higher-level and lower-level cortical fields happens very fast (Kral, 2013) and even within the auditory cortex, processing is characterized by complex forward, lateral and backward microcircuits (Atencio and Schreiner, 2010; Schreiner and Polley, 2014). Therefore, the origin of the categorical responses in AC is difficult to determine unless the response latencies and laminar structure are carefully investigated.

CROSSMODAL PLASTICITY—CONSIDERATIONS FOR FUTURE STUDIES

Considering that sound perception strongly relies on the integration of information represented across multiple cortical areas, simultaneous input from the other sensory modalities presents itself as a major source of influence on learning-induced plasticity of sound representations. In fact, there is compelling behavioral evidence that the human perceptual system integrates specific, event-relevant information across auditory and visual (McGurk and MacDonald, 1976) or auditory and tactile (Gick and Derrick, 2009) modalities and that mechanisms of multisensory integration can be shaped through experience (Wallace and Stein, 2007). Together, these two facts predict that visual or tactile contexts during learning have a major impact on perceptual reorganization of sound representations.

Promising insights are provided by behavioral studies showing that multimodal training designs are generally superior to unimodal training designs (Shams and Seitz, 2008). The beneficial effect of multisensory exposure during training may last beyond the training period itself reflected in increased performance after removal of the stimulus from one modality (for review, see Shams et al., 2011). This effect has been demonstrated even for brief training periods and arbitrary stimulus pairs (Ernst, 2007), promoting the view that short-term multisensory learning can lead to lasting reorganization of the processing networks (Kilian-Hütten et al., 2011a,b). Given the considerable evidence for response modulation of auditory neurons by simultaneous non-acoustic events and even crossmodal activation of the auditory cortex in absence of sound stimuli (Calvert et al., 1997; Foxe et al., 2002; Fu et al., 2003; Brosch et al., 2005; Kayser et al., 2005; Pekkola et al., 2005; Schürmann et al., 2006; Nordmark et al., 2012), it is likely that sound representations at the level of AC are also prone to influences from the visual or tactile modality. Animal electrophysiology has suggested different laminar profiles for tactile and visual pathways in the auditory cortex indicative for forward and backward directed input respectively (Schroeder and Foxe, 2002). Crucially, the quasi-laminar resolution achievable with state-of-art ultra-high field fMRI (Polimeni et al., 2010) provides new possibility to systematically investigate—in humans—the detailed neurophysiological basis underlying the influence of non-auditory input on sound perception and on learning induced plasticity in sound representations in the auditory cortex.

CONCLUSION

In recent years, the phenomenon of perceptual categorization has stimulated a tremendous amount of research on the neural representation of perceptual sound categories in animals and humans. Despite this large data pool, no clear answer could yet be found on where abstract sound categories are represented in the brain. Whereas animal research provides increasing evidence for complex processing abilities of early auditory areas, results from human studies tend to promote more hierarchical processing models in which categorical perception relies on higher order temporal and frontal regions. In this review, we discussed this apparent discrepancy and illustrated the potential pitfalls attached to research on categorical sound processing. Separating perceptual and acoustical processes possibly represents

the biggest challenge. In this respect, it is crucial to note that many "perceptual" effects, demonstrated in animal studies, did not manifest as changes in overall signal level. Recent research has shown that while these effects may remain inscrutable to univariate contrast analyses typically employed in human neuroimaging, modern analysis techniques—such as fMRI-decoding—is capable of unraveling perceptual processes in locally distributed activation patterns. It is also becoming increasingly evident that in order to grasp the full capacity of auditory processing in low-level auditory areas, it is necessary to consider its susceptibility to context and task, flexibly adapting its processing resources according to the environmental demands. In order to bring the advances from animal and human research closer together, future approaches on categorical sound representations in humans are likely to require an integrative combination of controlled stimulation designs, sensitive measurement techniques (e.g., high field fMRI) and advanced analysis techniques.

ACKNOWLEDGMENTS

This work was supported by Maastricht University, Tilburg University and the Netherlands Organization for Scientific Research (NWO; VICI grant 453-12-002 to Elia Formisano).

REFERENCES

Ahissar, M., and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* 8, 457–464. doi: 10.1016/j.tics.2004.08.011

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

Atencio, C. A., and Schreiner, C. E. (2010). Laminar diversity of dynamic sound processing in cat primary auditory cortex. *J. Neurophysiol.* 192–205. doi: 10.1152/jn.00624.2009

Bathellier, B., Ushakova, L., and Rumpel, S. (2012). Discrete neocortical dynamics predict behavioral categorization of sounds. *Neuron* 76, 435–449. doi: 10.1016/j.neuron.2012.07.008

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature* 403, 309–312. doi: 10.1038/35002078

Benders, T., Escudero, P., and Sjerps, M. (2010). The interrelaton between acoustic context effects and available response categories in speech sound categorization. *J. Acoust. Soc. Am.* 131, 3079–3087. doi: 10.1121/1.3688512

Benzeguiba, M., De Mori, R. Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: a review. *Speech Commun.* 49, 10–11. doi: 10.1016/j.specom.2007.02.006

Bidelman, G. M., Moreno, S., and Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage* 79, 201–212. doi: 10.1016/j.neuroimage.2013.04.093

Boettiger, C. A., and D'Esposito, M. (2005). Frontal networks for learning and executing arbitrary stimulus-response associations. *J. Neurosci.* 25, 2723–2732. doi: 10.1523/JNEUROSCI.3697-04.2005

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., and Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J. Neurosci.* 34, 4548–4557. doi: 10.1523/JNEUROSCI.4339-13.2014

Bonte, M., Valente, G., and Formisano, E. (2009). Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. *J. Neurosci.* 29, 1699–1706. doi: 10.1523/JNEUROSCI.3694-08.2009

Bornstein, M. H., Kessen, W., and Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *J. Exp. Psychol. Hum. Percept. Perform.* 2, 115–129. doi: 10.1037/0096-1523.2.1.115

Boynton, G. M., and Finney, E. M. (2003). Orientation-specific adaptation in human visual cortex. *J. Neurosci.* 23, 8781–8787.

Brechmann, A., and Scheich, H. (2005). Hemispheric shifts of sound representation in auditory cortex with conceptual listening. *Cereb. Cortex* 15, 578–587. doi: 10.1093/cercor/bhh159

Brosch, M., Selezneva, E., and Scheich, H. (2005). Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. *J. Neurosci.* 25, 6797–6806. doi: 10.1523/JNEUROSCI.1571-05.2005

Burns, E. M., and Ward, W. D. (1978). Categorical perception-phenomenon or epiphenomenon: evidence from experiments in the perception of melodic musical intervals. *J. Acoust. Soc. Am.* 63, 456–468. doi: 10.1121/1.381737

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

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

Caporale, N., and Dan, Y. (2008). Spike timing-dependent plasticity: a Hebbian learning rule. *Annu. Rev. Neurosci.* 31, 25–46. doi: 10.1146/annurev.neuro.31.060407.125639

Chechik, G., and Nelken, I. (2012). Auditory abstraction from spectro-temporal features to coding auditory entities. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18968–18973. doi: 10.1073/pnas.1111242109

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

De Baene, W., Ons, B., Wagemans, J., and Vogels, R. (2008). Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. *Learn. Mem.* 15, 717–727. doi: 10.1101/lm.1040508

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., and Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage* 24, 21–33. doi: 10.1016/j.neuroimage.2004.09.039

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

De Souza, A. C. S., Yehia, H. C., Sato, M., and Callan, D. (2013). Brain activity underlying auditory perceptual learning during short period training: simultaneous fMRI and EEG recording. *BMC Neurosci.* 14:8. doi: 10.1186/1471-2202-14-8

Eimas, P. D., Miller, J. L., and Jusczyk, P. W. (1987). "On infant speech perception and the acquisition of language," in *Categorical Perception.* The Groundwork of Cognition, ed S. Harnad (Cambridge, MA: Cambridge University Press), 161–195.

Engineer, C. T., Perez, C. A., Carraway, R. S., Chang, K. Q., Roland, J. L., and Kilgard, M. P. (2014). Speech training alters tone frequency tuning in rat primary auditory cortex. *Behav. Brain Res.* 258, 166–178. doi: 10.1016/j.bbr.2013.10.021

Engineer, C. T., Perez, C. A., Carraway, R. S., Chang, K. Q., Roland, J. L., Sloan, A. M., et al. (2013). Similarity of cortical activity patterns predicts generalization behavior. *PLoS ONE* 8:e78607. doi: 10.1371/journal.pone.0078607

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *J. Vis.* 7, 1–14. doi: 10.1167/7.5.7

Etcoff, N. L., and Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition* 44, 227–240. doi: 10.1016/0010-0277(92)90002-Y

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322, 970–973. doi: 10.1126/science.1164318

Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., et al. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *J. Neurophysiol.* 88, 540–543. doi: 10.1151/jn.00694.2001

Franklin, A., and Davies, I. R. L. (2004). New evidence for infant colour categories. *Br. J. Dev. Psychol.* 22, 349–377. doi: 10.1348/0261510041552738

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316. doi: 10.1126/science.291.5502.312

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.* 23, 5235–5246.

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

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Fu, K.-M. G., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., et al. (2003). Auditory cortical neurons respond to somatosensory stimulation. *J. Neurosci.* 23, 7510–7515.

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

Gick, B., and Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature* 462, 502–504. doi: 10.1038/nature08572

Golestani, N., and Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Goudbeek, M., Swingley, D., and Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1913–1933. doi: 10.1037/a0015781

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nat. Rev. Neurosci.* 5, 887–892. doi: 10.1038/nrn1538

Guenther, F. H., Husain, F. T., Cohen, M. A., and Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *J. Acoust. Soc. Am.* 106, 2900–2912. doi: 10.1121/1.428112

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., and Tourville, J. A. (2004). Representation of sound categories in auditory cortical maps. *J. Speech Lang. Hear. Res.* 47, 46–57. doi: 10.1044/1092-4388(2004/005)

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hear. Res.* 271, 133–146. doi: 10.1016/j.heares.2010.01.011

Harnad, S. (eds.). (1987). *Categorical Perception: The Groundwork of Cognition.* Cambridge: Cambridge University Press.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage* 62, 852–855. doi: 10.1016/j.neuroimage.2012.03.016

Haynes, J.-D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. doi: 10.1038/nrn1931

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Wiley.

Husain, F. T., Fromm, S. J., Pursley, R. H., Hosey, L., Braun, A., and Horwitz, B. (2006). Neural bases of categorization of simple speech and nonspeech sounds. *Hum. Brain Mapp.* 27, 636–651. doi: 10.1002/hbm.20207

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron* 53, 891–903. doi: 10.1016/j.neuron.2007.02.015

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron* 48, 373–384. doi: 10.1016/j.neuron.2005.09.018

Kilian-Hütten, N., Valente, G., Vroomen, J., and Formisano, E. (2011a). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *J. Neurosci.* 31, 1715–1720. doi: 10.1523/JNEUROSCI.4572-10.2011

Kilian-Hütten, N., Vroomen, J., and Formisano, E. (2011b). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *Neuroimage* 57, 1601–1607. doi: 10.1016/j.neuroimage.2011.05.043

Klein, M. E., and Zatorre, R. J. (2011). A role for the right superior temporal sulcus in categorical perception of musical chords. *Neuropsychologia* 49, 878–887. doi: 10.1016/j.neuropsychologia.2011.01.008

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

Köver, H., Gill, K., Tseng, Y.-T. L., and Bao, S. (2013). Perceptual and neuronal boundary learned from higher-order stimulus probabilities. *J. Neurosci.* 33, 3699–3705. doi: 10.1523/JNEUROSCI.3166-12.2013

Kral, A. (2013). Auditory critical periods: a review from system's perspective. *Neuroscience* 247, 117–133. doi: 10.1016/j.neuroscience.2013.05.021

Kraus, N., Skoe, E., Parbery-Clark, A., and Ashley, R. (2009). Experience-induced malleability in neural encoding of pitch, timbre, and timing. *Ann. N.Y. Acad. Sci.* 1169, 543–557. doi: 10.1111/j.1749-6632.2009.04549.x

Kuhl, P. K. (2000). A new view of language acquisition. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11850–11857. doi: 10.1073/pnas.97.22.11850

Kuhl, P. K., and Miller, J. D. (1975). Speech perception by the chinchilla: voiced-voiceless distrinction in alveolar plosive consonants. *Science* 190, 69–72. doi: 10.1126/science.1166301

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606–608. doi: 10.1126/science.1736364

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., and Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron* 77, 750–761. doi: 10.1016/j.neuron.2012.11.034

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Leech, R., Holt, L. L., Devlin, J. T., and Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J. Neurosci.* 29, 5234–5239. doi: 10.1523/JNEUROSCI.5758-08.2009

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., and Formisano, E. (2012). Learning of new sound categories shapes neural response patterns in human auditory cortex. *J. Neurosci.* 32, 13273–13280. doi: 10.1523/JNEUROSCI.0584-12.2012

Li, S., Mayhew, S. D., and Kourtzi, Z. (2009). Learning shapes the representation of behavioral choice in the human brain. *Neuron* 62, 441–452. doi: 10.1016/j.neuron.2009.03.016

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417

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

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Liebenthal, E., Sabri, M., Beardsley, S. A., Mangalathu-Arumana, J., and Desai, A. (2013). Neural dynamics of phonological processing in the dorsal auditory stream. *J. Neurosci.* 33, 15414–15424. doi: 10.1523/JNEUROSCI.1511-13.2013

Little, D. M., and Thulborn, K. R. (2005). Correlations of cortical activation and behavior during the application of newly learned categories. *Brain Res. Cogn. Brain Res.* 25, 33–47. doi: 10.1016/j.cogbrainres.2005.04.015

Logan, J. S., Lively, S. E., and Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *J. Acoust. Soc. Am.* 89, 874–886. doi: 10.1121/1.1894649

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

Minshew, N. J., Meyer, J., and Goldstein, G. (2002). Abstract reasoning in autism: a disassociation between concept formation and concept identification. *Neuropsychology* 16, 327–334. doi: 10.1037/0894-4105.16.3.327

Mizrahi, A., Shalev, A., and Nelken, I. (2014). Single neuron and population coding of natural sounds in auditory cortex. *Curr. Opin. Neurobiol.* 24, 103–110. doi: 10.1016/j.conb.2013.09.007

Murray, M. M., Camen, C., Gonzalez Andino, S. L., Bovet, P., and Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *J. Neurosci.* 26, 1293–1302. doi: 10.1523/JNEUROSCI.4511-05.2006

Myers, E. B., Blumstein, S. E., Walsh, E., and Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychol. Sci.* 20, 895–903. doi: 10.1111/j.1467-9280.2009.02380.x

Myers, E. B., and Swan, K. (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. *J. Cogn. Neurosci.* 24, 1695–1708. doi: 10.1162/jocn_a_00243

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Curr. Opin. Neurobiol.* 14, 474–480. doi: 10.1016/j.conb.2004.06.005

Niziolek, C. A., and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J. Neurosci.* 33, 12090–12098. doi: 10.1523/JNEUROSCI.1008-13.2013

Nordmark, P. F., Pruszynski, J. A., and Johansson, R. S. (2012). BOLD responses to tactile stimuli in visual and auditory cortex depend on the frequency fontent of stimulation. *J. Cogn. Neurosci.* 24, 2120–2134. doi: 10.1162/jocn_a_00261

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Nourski, K. V., and Brugge, J. F. (2011). Representation of temporal sound features in the human auditory cortex. *Rev. Neurosci.* 22, 187–203. doi: 10.1515/rns.2011.016

Ohl, F. W., Scheich, H., and Freeman, W. J. (2001). Change in pattern of ongoing cortical activity with auditory category learning. *Nature* 412, 733–736. doi: 10.1038/35089076

Olshausen, B. A., and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007

Ouimet, T., Foster, N. E. V., Tryfon, A., and Hyde, K. L. (2012). Auditory-musical processing in autism spectrum disorders: a review of behavioral and brain imaging studies. *Ann. N.Y. Acad. Sci.* 1252, 325–331. doi: 10.1111/j.1749-6632.2012.06453.x

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3T. *Neuroreport* 16, 125–128. doi: 10.1097/00001756-200502080-00010

Polimeni, J. R., Fischl, B., Greve, D. N., and Wald, L. L. (2010). Laminar analysis of 7T BOLD using an imposed spatial activation pattern in human V1. *Neuroimage* 52, 1334–1346. doi: 10.1016/j.neuroimage.2010.05.005

Polley, D. B., Steinberg, E. E., and Merzenich, M. M. (2006). Perceptual learning directs auditory cortical map reorganization through top-down influences. *J. Neurosci.* 26, 4970–4982. doi: 10.1523/JNEUROSCI.3771-05.2006

Prather, J. F., Nowicki, S., Anderson, R. C., Peters, S., and Mooney, R. (2009). Neural correlates of categorical perception in learned vocal communication. *Nat. Neurosci.* 12, 221–228. doi: 10.1038/nn.2246

Raizada, R. D., and Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron* 56, 726–740. doi: 10.1016/j.neuron.2007.11.001

Recanzone, G. H., Schreiner, C. E., and Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J. Neurosci.* 13, 87–103.

Scharinger, M., Henry, M. J., Erb, J., Meyer, L., and Obleser, J. (2013a). Thalamic and parietal brain morphology predicts auditory category learning. *Neuropsychologia* 53C, 75–83. doi: 10.1016/j.neuropsychologia.2013.09.012

Scharinger, M., Henry, M. J., and Obleser, J. (2013b). Prior experience with negative spectral correlations promotes information integration during auditory category learning. *Mem. Cogn.* 41, 752–768. doi: 10.3758/s13421-013-0294-9

Scheich, H., Brechmann, A., Brosch, M., Budinger, E., and Ohl, F. W. (2007). The cognitive auditory cortex: task-specificity of stimulus representations. *Hear. Res.* 229, 213–224. doi: 10.1016/j.heares.2007.01.025

Schreiner, C. E., and Polley, D. B. (2014). Auditory map plasticity: diversity in causes and consequences. *Curr. Opin. Neurobiol.* 24, 143–156. doi: 10.1016/j.conb.2013.11.009

Schroeder, C. E., and Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Res. Cogn. Brain Res.* 14, 187–198. doi: 10.1016/S0926-6410(02)00073-3

Schürmann, M., Caetano, G., Hlushchuk, Y., Jousmäki, V., and Hari, R. (2006). Touch activates human auditory cortex. *Neuroimage* 30, 1325–1331. doi: 10.1016/j.neuroimage.2005.11.020

Seitz, A. R., and Watanabe, T. (2003). Is subliminal learning really passive? *Nature* 422, 2003. doi: 10.1038/422036a

Shamma, S., and Fritz, J. (2014). Adaptive auditory computations. *Curr. Opin. Neurobiol.* 25C, 164–168. doi: 10.1016/j.conb.2014.01.011

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

Shams, L., Wozny, D. R., Kim, R., and Seitz, A. (2011). Influences of multisensory experience on subsequent unisensory processing. *Front. Psychol.* 2:264. doi: 10.3389/fpsyg.2011.00264

Sigala, N., and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415, 318–320. doi: 10.1038/415318a

Spierer, L., De Lucia, M., Bernasconi, F., Grivel, J., Bourquin, N. M., Clarke, S., et al. (2011). Learning-induced plasticity in human audition: objects, time, and space. *Hear. Res.* 271, 88–102. doi: 10.1016/j.heares.2010.03.086

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.* 19, 498–502. doi: 10.1016/j.cub.2009.01.066

Sussman, E., Winkler, I., Huotilainen, M., Ritter, W., and Näätänen, R. (2002). Top-down effects can modify the initially stimulus-driven auditory organization. *Brain Res. Cogn. Brain Res.* 13, 393–405. doi: 10.1016/S0926-6410(01)00131-8

Tsunada, J., Lee, J. H., and Cohen, Y. E. (2011). Representation of speech categories in the primate auditory cortex. *J. Neurophysiol.* 105, 2634–2646. doi: 10.1152/jn.00037.2011

Tsunada, J., Lee, J. H., and Cohen, Y. E. (2012). Differential representation of auditory categories between cell classes in primate auditory cortex. *J. Physiol.* 590, 3129–3139. doi: 10.1113/jphysiol.2012.232892

Tzounopoulos, T., and Kraus, N. (2009). Learning to encode timing: mechanisms of plasticity in the auditory brainstem. *Neuron* 62, 463–469. doi: 10.1016/j.neuron.2009.05.002

Verhoef, B. E., Kayaert, G., Franko, E., Vangeneugden, J., and Vogels, R. (2008). Stimulus similarity-contingent neural adaptation can be time and cortical area dependent. *J. Neurosci.* 28, 10631–10640. doi: 10.1523/JNEUROSCI.3333-08.2008

Wallace, M. T., and Stein, B. E. (2007). Early experience determines how the senses will interact. *J. Neurophysiol.* 97, 921–926. doi: 10.1152/jn.00497.2006

Watanabe, T., Náñez, J. E., and Sasaki, Y. (2001). Perceptual learning without perception. *Nature* 413, 844–848. doi: 10.1038/35101601

Wyttenbach, R. A., May, M. L., and Hoy, R. R. (1996). Categorical perception of sound frequency by crickets. *Science* 273, 1542–1544. doi: 10.1126/science.273.5281.1542

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

Received: 02 March 2014; accepted: 14 May 2014; published online: 03 June 2014.
*Citation: Ley A, Vroomen J and Formisano E (2014) How learning to abstract shapes neural sound representations. Front. Neurosci. **8**:132. doi: 10.3389/fnins.2014.00132*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 04 June 2014
doi: 10.3389/fnins.2014.00137

Simultaneous EEG-fMRI brain signatures of auditory cue utilization

Mathias Scharinger[1]*, Björn Herrmann[1], Till Nierhaus[2] and Jonas Obleser[1]

[1] *Max Planck Research Group "Auditory Cognition," Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[2] *Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Lee M. Miller, UC Davis, USA
Scott A. Beardsley, Marquette University, USA

****Correspondence:***
Mathias Scharinger, Department of Language and Literature, Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, 60322 Frankfurt am Main, Germany
e-mail: mscharinger@cbs.mpg.de

Optimal utilization of acoustic cues during auditory categorization is a vital skill, particularly when informative cues become occluded or degraded. Consequently, the acoustic environment requires flexible choosing and switching amongst available cues. The present study targets the brain functions underlying such changes in cue utilization. Participants performed a categorization task with immediate feedback on acoustic stimuli from two categories that varied in duration and spectral properties, while we simultaneously recorded Blood Oxygenation Level Dependent (BOLD) responses in fMRI and electroencephalograms (EEGs). In the first half of the experiment, categories could be best discriminated by spectral properties. Halfway through the experiment, spectral degradation rendered the stimulus duration the more informative cue. Behaviorally, degradation decreased the likelihood of utilizing spectral cues. Spectrally degrading the acoustic signal led to increased alpha power compared to nondegraded stimuli. The EEG-informed fMRI analyses revealed that alpha power correlated with BOLD changes in inferior parietal cortex and right posterior superior temporal gyrus (including planum temporale). In both areas, spectral degradation led to a weaker coupling of BOLD response to behavioral utilization of the spectral cue. These data provide converging evidence from behavioral modeling, electrophysiology, and hemodynamics that (a) increased alpha power mediates the inhibition of uninformative (here spectral) stimulus features, and that (b) the parietal attention network supports optimal cue utilization in auditory categorization. The results highlight the complex cortical processing of auditory categorization under realistic listening challenges.

Keywords: audition, categorization, cue weighting, spectro-temporal information, alpha suppression, attention

INTRODUCTION

The interpretation of acoustic signals is an essential human skill for goal-directed behavior and vocal communication. The core process underlying this skill—auditory categorization—has been shown to be highly flexible and adaptive, and allows, for instance, speaker recognition in a cocktail party situation (Zion Golumbic et al., 2013), or speech comprehension in noise (Nahum et al., 2008). In both cases, attention has to be directed to the most informative aspect of the acoustic signal (Hill and Miller, 2010).

Neurophysiological studies have suggested that the relative weighting of information during categorization (*information gain* or *cue weighting*, cf. Holt and Lotto, 2006) may be subserved by the interplay between excitatory and inhibitory mechanisms (Thut et al., 2006; Rihs et al., 2007; Weissman et al., 2009). One promising neurophysiological marker of functional inhibition processes are brain oscillations recorded using electroencephalography (EEG), predominantly in the alpha frequency range (8–13 Hz, Foxe et al., 1998; Foxe and Snyder, 2011; Weisz et al., 2011, 2013; Klimesch, 2012). Initially, alpha power had been interpreted as reflecting the degree to which primary cortical areas are in an "idling" mode (Adrian and Matthews, 1934; Niedermeyer and Silva, 2005). More recent studies on auditory comprehension, on the other hand, have shown that the processing of degraded speech stimuli is accompanied by relative decreases in alpha power suppression, i.e., relative increases in alpha power (Obleser and Weisz, 2012; Becker et al., 2013). One interpretation of this finding is that relative increases in alpha power index greater attention and working memory demands under degradation (Ronnberg et al., 2008; Wild et al., 2012). It has been further proposed that brain regions showing high alpha power undergo inhibition, which in turn allows enhanced processing of task-relevant information (Klimesch et al., 2007).

Brain areas underlying the processing and categorization of acoustic information have been identified by means of functional magnetic resonance imaging (fMRI). Previous studies have shown that the posterior part of the superior temporal gyrus (pSTG) is crucially involved in auditory categorization and discrimination (Hall et al., 2002; Guenther et al., 2004; Husain et al., 2006; Desai et al., 2008; Bermudez et al., 2009; Sharda and Singh, 2012). Importantly, in most of these studies, auditory categorization was also subserved by the planum temporale (PT) in the pSTG. The PT has recently received particular attention, because it does not only play a general role in auditory categorization (Griffiths and Warren, 2002; Husain et al., 2006; Obleser and Eisner, 2009) but also a more specific one with regard to the

processing of spectral information and pitch (Hall and Plack, 2009; Alho et al., 2014).

Furthermore, feature-selective attentional processes play a crucial role in categorization. Studies concerned with aspects of selective attention during categorization have mainly focused on the visual system (Yantis, 1993; Posner and Dehaene, 1994; Corbetta et al., 2000; Yantis, 2008). These studies identified the inferior parietal lobule (IPL) as an important, hub-like structure, being involved when participants focus attention on informative stimulus features (Shaywitz et al., 2001; Behrmann et al., 2004; Geng and Mangun, 2009; Salmi et al., 2009; Schultz and Lennert, 2009; Gillebert et al., 2012). Existing research on attention in audition has further provided evidence for the involvement of the parietal network (Rinne et al., 2007; Salmi et al., 2009; Hill and Miller, 2010; Henry et al., 2013). In addition, a recent structural imaging (voxel-based morphometry) study also highlighted the role of the IPL in categorization processes (Scharinger et al., 2014).

More recently, the possibility to combine recordings of EEG oscillatory activity and fMRI Blood Oxygenation Level Dependent (BOLD) activity has been explored in several imaging studies. Simultaneous EEG–fMRI recordings (Ritter and Villringer, 2006; Sadaghiani et al., 2010, 2012) suggest that alpha power can be negatively (Goldman et al., 2002; Laufs et al., 2003; Ritter and Villringer, 2006) or positively (Moosmann et al., 2003; Liu et al., 2012) correlated with brain metabolism, depending on the brain regions these correlations are observed in. However, multi-modal neuroimaging evidence on auditory cue weighting during categorization has been essentially absent. Most studies concerned with a functional coupling of alpha power and BOLD signal in selective attention tasks compared the processing of task-relevant information with the processing of task-irrelevant distractor information (e.g., Scheeringa et al., 2012).

It is thus less clear how multiple, potentially competing cues provided by the same acoustic stimulus, will be reflected in alpha-tuned functional processes and concomitant BOLD change. To this end, we designed two stimulus sets for auditory categorization. In the first stimulus set, categorization could be based on spectral properties or physical duration, with spectral properties being more informative. In the second stimulus set, sound duration became the more informative cue, while spectral properties could still be used for categorization. Using combined EEG/fMRI, we asked (a) whether auditory categorization yields a behavioral preference for the most informative stimulus cue in each condition; (b) which brain areas support change in cue utilization, (c) whether alpha power shows relative increases under degradation and (d) whether alpha power correlates with BOLD in brain areas dedicated to the processing of acoustic cues.

MATERIALS AND METHODS

PARTICIPANTS

Sixteen healthy volunteers were recruited from the participant database of the Max Planck Institute for Human Cognitive and Brain Sciences (7 females, age range 20–29 years, age 25 ± 2.7 years mean $\pm$ standard deviation). They were all right-handed, native speakers of German with no self-reported hearing impairments or neurological disorders. Due to technical problems with EEG acquisition in the magnetic resonance (MR) scanner, we had to exclude one participant from further analyses. Participants gave written informed consent and received financial compensation for their participation. All procedures followed the guidelines of the local ethics committee (University of Leipzig) and were in accordance with the Declaration of Helsinki.

STIMULI

Stimuli were based on spectral and durational modifications of an inharmonic base signal. This base signal was constructed by adding 16 exponentially spaced sinusoids (ratio between successive components: 1.15) to the lowest sinusoid component frequency of 500 Hz (Goudbeek et al., 2009; Scharinger et al., 2014). We modified the spectral properties of individual sounds by applying a band-pass filter with a single frequency peak, using a second order infinite impulse response (IIR) filter with a bandwidth corresponding to a fifth of its frequency peak. The term "spectral peak" is henceforth used to refer to the filters' center frequency, which also describes the resulting spectral properties. Duration modifications were based on differences in the length of the sounds.

Individual members of category distributions, arbitrarily labeled "A" and "B," varied on the basis of spectral peak and duration: For individual sounds of each category, spectral filter frequencies and durations were randomly drawn from bivariate normal distributions. These distributions, with equal standard deviations, σ, differed in their means, μ, between the two categories, A and B (**Table 1**). Thus, each individual sound was characterized by the two dimensions, duration and spectral peak, with means of duration and spectral peak differing between the two category distributions. Each category distribution consisted of 1000 sound exemplars from which a random sample was drawn for each participant in the experiment. Following Smits et al. (2006), we converted spectral peak frequency and duration to scales that allowed for psychoacoustic comparability. Consequently, frequencies were converted to the equivalent rectangular bandwidth (ERB) scale that approximates the bandwidths of the auditory filters in human hearing (Glasberg and Moore, 1990), and durations were converted to a logarithmic scale (DUR; cf. Smits et al., 2006). **Table 1** illustrates the means (spectral peak and durations) of the category distributions in psychophysical and physical units.

In the first half of the experiment (*nondegraded condition*), the two stimulus distributions did not overlap in their spectral peak, but $\frac{1}{3}$ of the sounds in category A and B overlapped in duration (**Figure 1A** top). This set-up aimed at biasing participants to focus on spectral cues while sound duration may serve as secondary cue. In the second half of the experiment (*degraded condition*), spectral cues were modified by applying four-band noise vocoding to the original stimulus distributions (Drullman et al., 1994; Shannon et al., 1995). Noise vocoding was done by dividing the original signal into four frequency bands, extracting the amplitude envelope from each band and reapplying it to bandpass-filtered noise carriers with matched cut-off frequencies. Envelopes were extracted using a zero-phase, 4th-order Butterworth low-pass filter; the low-pass filter cutoff was set at 256 Hz. Scaling for equal root mean square (RMS) energy

Table 1 | Means and standard deviations (in parentheses) of spectral peak and duration distributions for stimulus categories A and B in the nondegraded and degraded conditions (psychophysical and physical units).

Stimulus category	Nondegraded		Degraded	
	A	B	A	B
Spectral peak (ERB)	20.00 (0.31)	17.00 (0.31)	16.80 (0.31)	15.50 (0.31)
Spectral peak (Hz)	1739 (8)	1196 (8)	1166 (8)	984 (8)
Duration (DUR)	47.70 (1.31)	52.53 (1.31)	47.70 (1.31)	52.53 (1.31)
Duration (ms)	118 (1.14)	191 (1.14)	118 (1.14)	191 (1.14)

was performed channel-wise for each channel envelope (Rosen et al., 1999; Erb et al., 2012). We chose four-band noise vocoding because it offers a well-established reduction of spectrally-based intelligibility (cf. Scott et al., 2006; Obleser and Kotz, 2010; Obleser et al., 2012), thereby ensuring comparability to studies on alpha power suppression in speech, while simultaneously being an ecologically valid modification by simulating effects of cochlear implants (Poissant et al., 2006).

Noise vocoding led to a smearing of spectral detail, while amplitude envelope features and original stimulus duration remained unaffected (**Figure 1A**, bottom). Thus, as demonstrated before (Scharinger et al., 2014), we aimed at inducing a change in acoustic cue utilization, from spectral peak in the first (nondegraded) condition, to stimulus duration in the second (degraded) condition of the experiment. The stimulus degradation in the second half of the experiment therefore targeted the spectral properties (i.e., spectral peak, but also affected other spectral features such as harmonicity). Thus, degradation of the initially informative spectral cue ought to decrease participants' reliance on that cue and prompt a relatively increased reliance on the duration cue.

All stimuli were normalized for equal root-mean-square intensity and presented at ∼60 dB SPL. Onset and offset ramps (5 ms) ensured that acoustic artifacts were minimized.

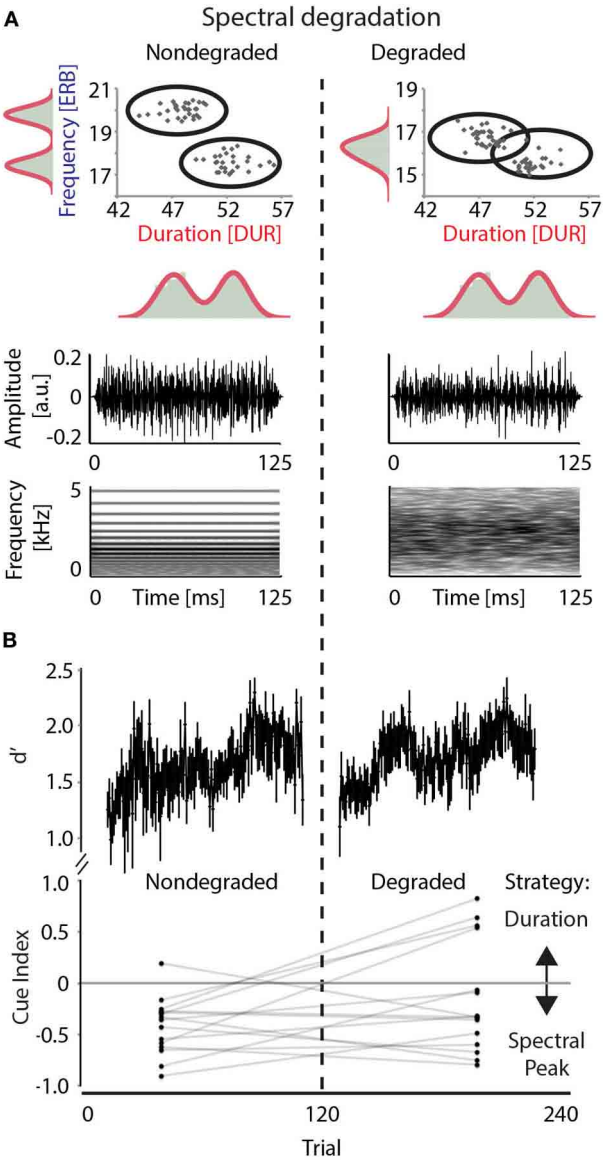


FIGURE 1 | Stimulus characteristics and behavioral results. (A) *Top*: Complex sounds differing in spectral peak (expressed in ERB; y-axis) and duration (expressed in DUR; x-axis). Distributions are indicated by ellipses, with black dots illustrating distributions for a representative participant. *Bottom*: Stimulus wave form and spectrogram illustrate the complex structure of sounds in the nondegraded condition (left) and the spectral smearing as a result of vocoding in the degraded condition (right). Duration and amplitude envelope were unaffected by degradation. **(B)** Results of behavioral discrimination. *Top*: Perceptual sensitivity (d') over time, obtained from sliding windows over nondegraded and degraded trials per participant (window size = 20 trials, step size = 1 trial). *Bottom*: Comparison between by-subject cue indices in the nondegraded and degraded conditions. Mean cue index values for the nondegraded and the degraded conditions are connected for each participant.

EXPERIMENTAL PROCEDURE

Participants were first familiarized with the categorization task in the scanner and had to complete a short practice run consisting of 20 sounds (10 from category A and 10 from category B) that did not occur in the main experiment. The subsequent main experiment was arranged in four runs: Two initial runs with nondegraded sounds, and two subsequent runs with spectrally degraded sounds (**Figure 1A**, top). In each run, 60 sound exemplars, randomly drawn from categories A and B with equal probability, were presented in a sparse imaging design in the MR scanner (Hall et al., 1999). The sparse design was chosen in order to guarantee that stimuli could be presented during silent periods in-between the acquisition of echo-planar images (EPI). At the same time, this design reduced contamination of the EEG signal by gradient switches during volume acquisition.

On each trial, one acoustic stimulus was presented on average 2 s after the offset of a preceding EPI sequence (±500 ms). Subsequently, a visual response prompt (green traffic light) was presented on a screen which participants viewed through a mirror 3 s after stimulus onset. Participants were then required to indicate whether the presented sound belonged to category A or category B by pressing one of two keys on a button

box. Button assignment was counterbalanced across participants. Following the response, participants received corrective feedback (*Correct/Incorrect*), which was displayed for 1 s in the middle of the screen. Five seconds after the onset of an acoustic stimulus, a subsequent EPI volume (acquisition time TA = 2 s) was acquired, such that the BOLD peak would best capture stimulus processing. At random positions within each run, 15 silent trials (=20% of all trials) without required responses served as baseline. The duration of the entire experiment with short breaks between runs was 50 min.

ACQUISITION AND PRE-PROCESSING OF EEG DATA

The continuous EEG was recorded inside the MR-scanner from 31 Ag–AgCl electrodes mounted on an elastic cap according to the 10–20 standard system (EasyCap-MR, Brain Products, Munich, Germany). The electrocardiogram (ECG) was registered with an additional electrode on the sternum. EEG signals were amplified with an MR-conform 32-channel amplifier (BrainAmp MR; Brain Products, Munich, Germany) that did not get saturated by MR activity. Signals were recorded at a sampling frequency of 5000 Hz and a resolution of 16 bits, referenced against FCz, using the BrainVision Recorder Software (Brain Products, Munich, Germany). The ground electrode was positioned between Fz and FPz. All impedances were kept below 5 kΩ.

Since we used a sparse imaging design with stimuli being presented in-between two consecutive volume acquisitions, gradient artifact removal from the EEG was not necessary (cf. Herrmann and Debener, 2008; Huster et al., 2012). For preprocessing, a finite impulse response (FIR) 100 Hz low-pass filter (389 points, Hamming window) and a 1.7 Hz high-pass filter (4901 points, Hann window, corresponding to a cut-off period of 1/1.7 Hz = 588 ms) was applied to the raw data. Note that filter settings were chosen such that smearing of gradient artifacts into time windows of interest were prohibited. Subsequently, filtered EEG data were down-sampled to 500 Hz and subjected to an independent components analysis (ICA) for artifact correction, using the routines provided by EEGLab (Delorme and Makeig, 2004) and *fieldtrip* (Oostenveld et al., 2011) within MATLAB 7.9 (MathWorks, Natick, MA). Note that the ECG channel was removed prior to ICA analysis. ICAs were calculated on 3-s epochs, with 1 s before and 2 s after stimulus onset. The separation of ICA components (total: 29) representing artifacts from those representing physiological EEG activity was done by visual inspection of the components' time-courses, topographies, and frequency spectra (cf. Debener et al., 2010), using custom-made fieldtrip scripts. Components either showing similar dynamics as the ECG channel or resembling electrooculogram activity as illustrated in Debener et al. (2010) were considered artifacts. Note that it has been observed that ICA-based correction of cardio-ballistic artifacts performs better than standard artifact subtraction methods (Debener et al., 2007; Jann et al., 2009). On average, 7 components were therefore excluded (range: 5–9) by using the ICA-based artifact removal within fieldtrip (Oostenveld et al., 2011).

We furthermore identified bad EEG channels after artifact removal as channels exceeding a threshold of 150 μV in more than 50% of all trials per participant. Bad channels (of which no participant showed more than 1) were interpolated by using signal information from the average of 4–5 neighboring channels (depending on channel location).

In addition to EEG recordings inside the MR-scanner, we tested 18 different participants (9 females, mean age 25, range 20–31 years) outside the scanner. Presenting pre-recorded EPI sounds at times the scanner would have operated simulated the scanner noise. For this control group, the EEG was obtained from 64 Ag-AgCl-electrodes (58 scalp electrodes, 2 mastoids, 2 electrodes for horizontal and 2 for vertical electrooculograms) on a Brain Vision EEG system (amplifier: BrainAmp, cap: BrainCap, Brain Products, Munich, Germany), arranged according to the extended 10/20 system, (Oostenveld and Praamstra, 2001). Otherwise, stimulus presentation, EEG pre-processing and analyses were identical to the procedures described here. However, due to a technical problem with one participant, and more than 30% ICA-artifact components in two further participants, the resulting participant number of the control experiment was 15. This experiment served the purpose of testing the validity of the recordings obtained inside the scanner. Note, however, that overall magnitude differences should not be compared between the experiments inside and outside the scanner, due to different recording equipment.

ACQUISITION AND PRE-PROCESSING OF fMRI DATA

Functional MRI data were recorded with a Siemens VERIO 3.0-T MRI scanner equipped with a 12-channel head coil, while participants performed the categorization task in supine position inside the scanner. Acoustic stimuli were transmitted through MR-compatible headphones (mr confon GmbH, Magdeburg, Germany). In-ear hearing protection (Hearsafe Technologies GmbH, Cologne, Germany) reduced scanner noise by approximately 16 dB.

Seventy-five whole-brain EPI volumes (30 axial slices, thickness = 3 mm, gap = 1 mm) in each of the 4 runs were collected every 9 s (TA = 2 s; TE = 30 ms; flip angle = 90°; field of view = 192 × 192 mm; voxel size = 3 × 3× 4 mm). High-resolution, 3D MP-RAGE T1-weighted scans were used for localization and co-registration (acquired on a 3T Siemens TIM Trio scanner with a 12-channel head coil 29 months prior to the experiment, with the parameters: sagittal slices = 176, repetition time = 1300 ms, TE = 3.46 ms, flip angle = 10°, acquisition matrix = 256 × 240, voxel size = 1 × 1× 1 mm). Voxel-displacement-maps for distortion correction (Jezzard and Balaban, 1995; Hutton et al., 2002) were calculated on the basis of field maps (30 axial slices, thickness = 3 mm, gap = 1 mm, repetition time = 488 ms, TE1 = 4.92 ms, TE2 = 7.38 ms, flip angle = 60°, field of view = 192 × 192 mm, voxel size = 3 × 3× 3 mm).

Functional (T2*-weighted) and structural (T1-weighted) images were processed using Statistical Parametric Mapping (SPM8; Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College of London). Functional images were first realigned using the 6-parameter affine transformation in translational (x, y, and z) and rotational (pitch, roll, and yaw) directions to reduce individual movement artifacts (Ashburner and Good, 2003). Subsequently, a mean image of each run-based image series was used to estimate unwarping

parameters, and voxel-displacement-maps were used for correcting magnetic field deformations (Jezzard and Balaban, 1995; Hutton et al., 2002). Participants' structural images were manually pre-aligned to a standardized EPI template (Ashburner and Friston, 2004) in MNI space, improving co-registration and normalization accuracy. Next, functional images were co-registered to the corresponding participants' structural images and normalized to MNI space. Functional images were then smoothed using an 8-mm full-width half-maximum Gaussian kernel and subsequently used for first-level general linear model (GLM) analyses.

ANALYSIS OF BEHAVIORAL DATA

Our behavioral dependent measures were *overall performance* and *cue utilization*. Overall performance was estimated by d', a measure of perceptual sensitivity that is independent of response bias. Perceptual sensitivity, d', was calculated from proportions of hits and false alarms according to a one-interval design (Macmillan and Creelman, 2005), where hits were defined as "category-A" responses to category-A stimuli, and false alarms were defined as "category-A" responses to category-B stimuli. Perceptual sensitivity was calculated separately for each experimental run (2 nondegraded, 2 degraded runs). In order to visualize performance over time, we additionally calculated d' values in sliding windows (size: 20 trials, step size: 1 trials), separately for the nondegraded and the degraded condition, and with the exclusion of null trials.

The measure of *cue index* quantified individual participants' cue utilization (spectral peak vs. physical duration) in the following way: First, for each condition, the likelihood of a category-A response was predicted from the stimulus' physical properties, spectral peak and duration, by means of logistic regressions. The slope of the regressions function, expressed by absolute β, indicated the degree to which the corresponding physical stimulus property influenced the categorical response ($\beta_{\text{spectral peak}}$; β_{duration}; Goudbeek et al., 2009; Scharinger et al., 2013). Note that $\beta_{\text{spectral peak}}$ and β_{duration} were estimated simultaneously. Second, the normalized difference between these β values (*cue index*) indicated participants' preference to rely on spectral peak (negative values according) or on duration (positive values).

$$\text{Cue index} = \frac{\beta_{\text{duration}} - \beta_{\text{spectral peak}}}{\beta_{\text{duration}} + \beta_{\text{spectral peak}}}$$

ANALYSIS OF EEG DATA

For the analysis of the event-related potentials (ERPs), single-trial EEG epochs were first re-referenced to linked mastoids (approximated by channels Tp9 and Tp10). Subsequently, epochs were filtered with a 20-Hz Butterworth low-pass filter and re-defined to include a pre-stimulus interval of 500 ms and a post-onset interval of 1500 ms. Baseline correction was applied by subtracting the mean amplitude of the −500 to 0 ms baseline interval from the epoch. Single-trials were averaged separately for the nondegraded and the degraded condition. Auditory N1 components (Näätänen and Picton, 1987) were identified by visual inspection in a time window between 100 and 150 ms post onset. Averaged amplitudes for Cz within the N1 time-window were compared between conditions (nondegraded, degraded) by means of dependent-samples t-tests.

For time-frequency analyses, re-referenced EEG-data were down-sampled to 125 Hz and then decomposed with a Morlet wavelets analysis (Bertrand and Pantev, 1994), centered on windows that slid in steps of 10 ms along the temporal dimension (−1 to 2 s). In the spectral dimension, we used 1-Hz bins from 1 to 30 Hz. Wavelet widths ranged from 1 to 8 cycles, equally spaced over the 30 frequency bins. Time-frequency analyses were done separately for nondegraded and degraded trials. Mean power values of a pre-stimulus baseline interval (−500 to −50 ms) were subtracted from the epoch. A time-frequency region of interest (ROI) was chosen according to the typical alpha-band interval (7–11 Hz) and according to epochs that previously showed the suppression effect in speech (400–700 ms post onset, e.g., Obleser and Weisz, 2012; Becker et al., 2013). A consistent and symmetric posterior electrode selection for subsequent EEG/fMRI correlations was based on electrodes where alpha power was strongest in above-mentioned ROI (within the nondegraded condition). These electrodes were: CP1, CP2, P7, P3, Pz, P4, P8, POz, O1, Oz, and O2. Averaged power values in the alpha ROI was compared between conditions by means of dependent-samples t-tests.

ANALYSIS OF fMRI DATA

Activated voxels were identified using the GLM approach (Friston, 2004). At the first level, a GLM was estimated for each participant with a first-order finite impulse response (FIR; window = 2 s) and a high-pass filter with a cut-off of 128 s, representing standard settings for sparse imaging designs (cf. Peelle et al., 2010). The design matrix included regressors for *sound trials* (corresponding to volumes following sound representations), the mean-centered single-trial parametric modulator *alpha power* (obtained from the ROI defined above), and *silent trials* (corresponding to volumes following null trials). Experimental runs were included as regressors of no interest (one for each run). Six additional regressors of no-interest accounted for the realignment-induced spatial deformations of the EPI volumes.

Resulting beta-maps were restricted to gray- and white matter. This information was obtained from group-averages based on individual T1-weighted scans. On the first level, the following contrasts were calculated (separately for nondegraded and degraded conditions): *sound trials* against implicit baseline and parametric modulator *alpha power* against implicit baseline. Furthermore, we calculated the contrasts nondegraded > degraded and degraded > nondegraded.

On the second level (group level), all contrasts were compared against zero using one-sample t-tests. Additionally, for each condition (nondegraded, degraded), sound-trial contrasts (against implicit baseline) from the first level were correlated with cue index using linear regression. Differences between nondegraded and degraded conditions in Cue index/BOLD correlation were assessed by testing the slopes of the linear regressions against each other using a dependent samples t-test.

For statistical thresholding of second-level activations, we used a threshold of $p < 0.005$ combined with a cluster extent of 15

voxels that corresponds to a whole-brain significance level of $p < 0.05$, as determined from a MATLAB-implemented Monte Carlo simulation (Slotnick et al., 2003; Erb et al., 2013).

In order to visualize BOLD modulation differences across conditions, ROIs of 10 mm radii were defined using the SPM toolbox MarsBaR (Brett et al., 2002). They were centered on the peak coordinates of significant clusters identified in the whole-brain analyses. For these regions, mean regression beta values were estimated for each participant. Note that no additional tests were conducted for these regions to avoid statistical circularity. Determination of anatomical locations was based on the Automated Anatomical Labeling Atlas (AAL; Tzourio-Mazoyer et al., 2002), and PT localization followed Westbury et al. (1999).

RESULTS

BEHAVIORAL DATA

Participants performed above chance as indicated by d' values significantly greater than zero [mean $d' = 1.51$, $SD = 0.43$; $t_{(14)} = 19.19$, $p < 0.01$]. Participants' performance was characterized by a considerable improvement over the first twenty trials, as estimated from sliding-window averages of d'-values (window size: 20 trials, step size: 1 trial, **Figure 1B** top). After degradation was introduced, performance dropped to the initial level, but quickly regained a stable plateau and did not differ overall from the nondegraded condition [nondegraded vs. degraded $t_{(14)} = 1.00$, $p = 0.32$].

Cue indices marginally differed between conditions [$t_{(14)} = 1.94$, $p = 0.07$], with more negative values for the nondegraded than the degraded condition. This means that the tendency of utilizing spectral cues (i.e., a negative cue index) in the nondegraded condition decreased in the degraded condition (i.e., a positive-going cue index). However, a spectral strategy was never entirely given up, as judged from overall still negative cue indices in the degraded condition (**Figure 1B**, bottom).

EEG DATA

The N1 (100–150 ms) of the ERP showed a typical central/midline topography (inside and outside the scanner). N1 mean amplitude marginally differed between the nondegraded and the degraded condition [$t_{(14)} = 1.9$, $p = 0.08$], with more negative values in the nondegraded than in the degraded condition. This effect reached significance outside the scanner [$t_{(14)} = 7.89$, $p < 0.01$; **Figure 2A**].

Alpha power (7–11 Hz) around 400–700 ms showed a central-posterior distribution and also differed significantly between conditions, with relatively higher alpha power for the degraded than for the nondegraded condition [$t_{(14)} = 2.06$, $p = 0.04$ **Figure 2B**]. Again, this effect also held for the control experiment outside the scanner [$t_{(14)} = 2.56$, $p = 0.03$; **Figure 2C**].

In order to assess the covariation of alpha power and cue index, we calculated correlations between mean alpha power and mean cue index per participant, and in addition, separately for the nondegraded and degraded condition. Overall, mean alpha power and mean cue index did not correlate significantly [$r = 0.28$, $t_{(14)} = 1.07$, $p = 0.30$]. This held both within the nondegraded [$r = 0.23$, $t_{(14)} = 0.85$, $p = 0.41$] and the degraded condition [$r = 0.16$, $t_{(14)} = 0.60$, $p = 0.56$].

fMRI DATA

Overall auditory categorization network in parietal and temporal areas

Results from group-level whole-brain analyses showed that the categorization of nondegraded and degraded sounds (compared to baseline) lead to activations in extensive bilateral temporo-parietal clusters, with peaks in inferior parietal lobule and post-central gyrus (see **Figure 3**). Furthermore, peaks in precentral and cingulate cortex were predominantly seen for nondegraded sounds, while degraded sounds showed activations in pSTG, PT, and Heschl's gyrus. Both conditions also revealed substantial activations in middle frontal gyrus (MFG), inferior frontal gyrus (IFG), and in the dorsal medial nucleus of left Thalamus.

More activation for degraded than for nondegraded sounds was found in right IFG (extending into the insula), left and right pSTG (including parts of PT, i.e., gray matter with a likelihood of 25–45% being in PT according to Westbury et al., 1999), as well as right STG (extending into the insula). A detailed overview of the clusters is provided in **Table 2**.

Alpha power covaries with BOLD activity in pSTG, PT, and IFG

Group-level whole-brain analyses showed that single-trial alpha power correlated positively with BOLD only in the degraded condition. Here, alpha power/BOLD correlations occurred in two clusters in IFG (comprising pars triangularis and ventral orbitofrontal cortex), in one cluster located in right pSTG (with 25–45% probability of being in PT), and in one cluster in right angular gyrus. In the nondegraded condition, alpha power/BOLD correlations did not survive the statistical threshold.

Stronger modulations of BOLD by alpha power could be observed in the orbital part of right IFG, as well as in bilateral pSTG, again comprising parts of the PT (with 25–45% probability according to Westbury et al., 1999; cf. **Table 3** and **Figure 4A**).

Cue index modulates BOLD activity in parietal attention and temporal auditory network

Group-level whole-brain regression analyses using the cue index showed positive correlations with BOLD in right MFG (anterior prefrontal cortex) only in the degraded condition. Here, a reduction of using spectral cues corresponded to an increased BOLD signal in anterior prefrontal cortex. By contrast, cue index/BOLD correlations in the nondegraded condition did not survive the statistical threshold.

Furthermore, positive cue index/BOLD correlations were stronger in the degraded than in the nondegraded condition in right dorso-lateral prefrontal cortex (covering parts of pars triangularis and pars opercularis), left pSTG/pSTS (extending into PT), left posterior MTG (involving parts in occipito-temporal cortex), right (ventral) IPL (involving parts of supramarginal gyrus and extending rostrally into postcentral gyrus; cf. **Table 3** and **Figure 4B**).

DISCUSSION

The two most important findings of this multimodal brain imaging study on auditory categorization are the following: First, auditory categorization of degraded stimuli yielded decreases in alpha power suppression (i.e., relative alpha power increases),

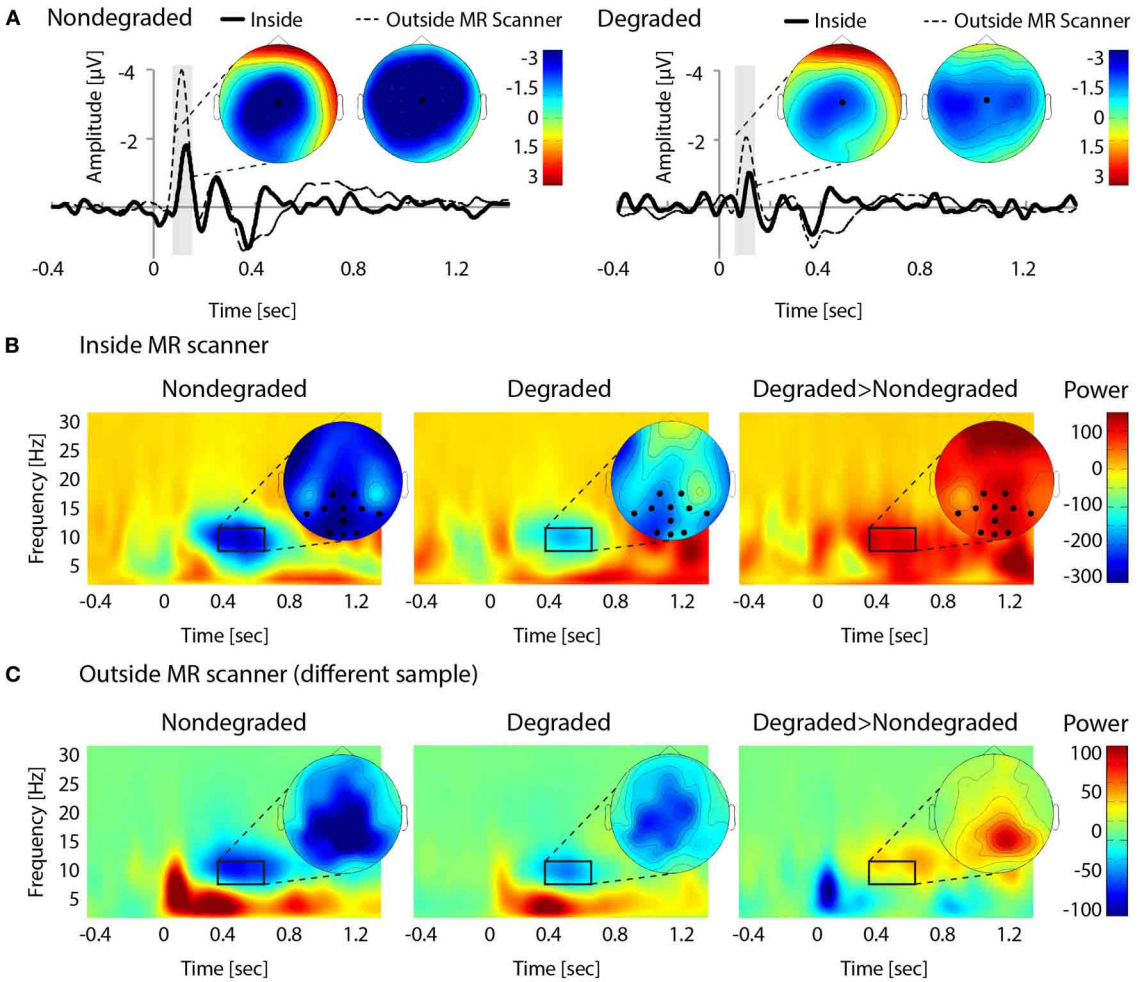


FIGURE 2 | EEG results. (A) Grand-average of evoked responses in the nondegraded (left) and degraded (right) condition. ERP-differences between conditions were seen for the N1, with a central/midline distribution (100–150 ms, indicated by gray bars). **(B)** Averaged time-frequency representations for the nondegraded (left) and degraded (middle) condition, and difference between averages (degraded > nondegraded; right). The strongest effect of alpha suppression (compared to baseline) occurred at central-posterior electrodes (selection marked with black dots; 400–700 ms, 7–11 Hz), where it also significantly differed between conditions. **(C)** Averaged time-frequency representations from the control experiment outside the MR scanner (nondegraded: left, degraded: middle, difference: right). Differences and topographies are comparable to within-scanner recordings. Note that overall magnitude differences should not be compared between the experiments inside and outside the MR scanner, due to different recording equipment.

which correlated with increased activation in right PT and IFG. Second, even though the behavioral measure of cue utilization only marginally differed between conditions, less reliance on spectral cues under sound degradation corresponded to increased activation in left PT and right IPL. In the subsequent sections, these findings will be discussed in more detail.

ENHANCED ALPHA POWER DURING DEGRADED SPEECH PROCESSING

In the current study, categorizing spectrally degraded sounds was accompanied by an attenuation of alpha power suppression. That is, relatively stronger alpha power was observed for the categorization of degraded as compared to nondegraded sounds. This reduction in alpha power suppression (relative to a pre-stimulus baseline) has previously been observed in comparing spectrally degraded speech stimuli to their nondegraded (intelligible) counter-parts (Obleser and Weisz, 2012; Becker et al., 2013). The current data thus extend previous findings by showing that increased alpha power under degradation is not restricted to speech material, but may reflect a more general process that has been interpreted before as enhanced "functional inhibition" (Jensen and Mazaheri, 2010), increased "idling" (Adrian and Matthews, 1934), or a more "active processing state" (Palva and Palva, 2011).

A parsimonious interpretation of this effect relates to the functional inhibition hypothesis of increased alpha power (e.g., Jensen and Mazaheri, 2010). According to this approach, alpha power shows a relative decrease in areas subserving the processing of to-be-attended information (Thut et al., 2006), while it increases in areas subserving the processing of to-be-ignored information (Rihs et al., 2007). Thereby, alpha power dynamics instate a

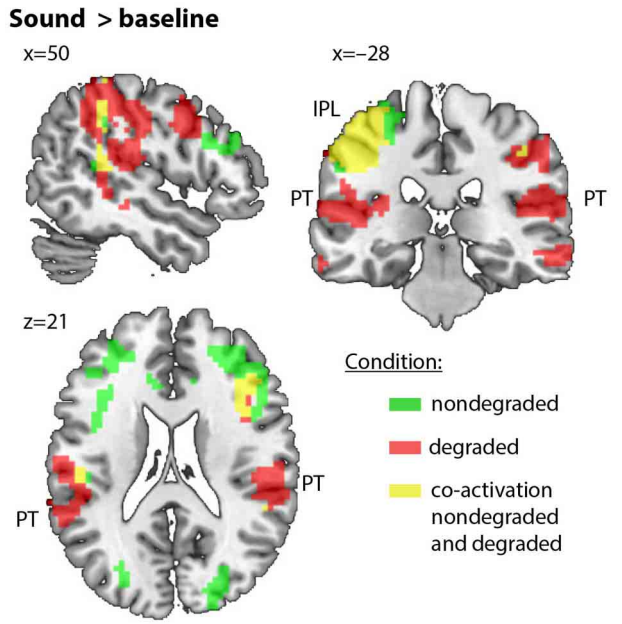


FIGURE 3 | Regions of the *sounds> baseline* contrast in nondegraded (green) and degraded (red) condition (co-activation in nondegraded and degraded condition: yellow). Slices focus on the temporo-parietal network. Note that overlap/co-activation is shown as illustrative means and is not based on statistical measures.

Table 2 | Significant clusters obtained from whole-brain analyses ($p < 0.005$, extent threshold = 15) for the contrasts sounds > baseline in each condition, and the contrast degraded sounds > nondegraded sounds.

Contrast	Area	Coordinates	*Z*	Extent (voxels)
Nondegraded sounds > baseline	l. IPL/BA40	−39, −13, 61	4.95	2659
	r. IFG/BA46	45, 38, 31	4.4	539
	r. IPL/SMG	42, −34, 46	4.28	470
	r. Cereb/Culmen	21, −55, −26	4.2	194
	r. Cereb/Culmen	3, −61, −32	4.18	170
	l. Thalamus	−6, −19, 7	3.87	113
	r. Cuneus	18, −91, 1	3.75	103
	l. Insula/BA13	−30, 14, 1	3.66	72
	r. Insula/BA13	30, 20, −2	3.65	61
	r. ITG/BA20	57, −46, −17	3.6	37
	l. Insula/BA13	−27, 26, −5	3.55	21
	l. Occ./BA17	−15, −91, 1	3.49	27
	l. MFG/BA10	−24, 59, −8	3.46	80
	l. pSTG/PT	−48, −46, 7	3.42	30
	r. pSTG/PT	51, −40, 13	3.17	35
Degraded sounds > baseline	l. Postcentral/IPL	−51, −22, 46	5.61	2074
	r. IPL/BA40	39, −43, 58	4.82	1563
	r. Cingulate/BA32	3, 11, 55	4.77	631
	r. Precentral/BA6	48, 5, 40	4.29	354
	l. Cuneus/BA18	−18, −100, 1	4.24	512
	r. MFG/BA11	21, 47, −11	4.24	15
	r. MFG/BA10	36, 50, 10	4	84
	r. IFG/BA47	30, 29, −2	3.7	70
	l. MFG/BA10	−33, 41, 4	3.7	79
	l. MTG/BA21	−63, −31, −14	3.64	41
	l. Thalamus	−12, −19, 10	3.47	63
	l. Insula/BA13	−30, 32, 7	3.32	75
	l. MFG/BA10	−27, 32, 25	3.2	19
	r. Cereb./Culmen	15, −52, −23	3.17	21
Degraded > Nondegraded	r. IFG/Insula	33, 14, −17	3.9	43
	l. pSTG/PT	−51, −37, 10	3.41	16
	r. STG	48, −4, −8	3.3	31
	r. pSTG/PT	54, −25, 19	3.2	30

Abbreviations are explained in the text. Coordinates are given in Montreal Neurological Institute (MNI) space.

gain mechanism for neural information processing (Jokisch and Jensen, 2007; Kerlin et al., 2010). While the functional role of alpha oscillations in auditory processing and categorization has been examined much less often and only recently (Weisz et al., 2011, 2013; Obleser and Weisz, 2012; Obleser et al., 2012; Becker et al., 2013), the interpretations provided by these previous studies are in line with the functional inhibition hypothesis. For instance, it has been observed that alpha power suppression correlates with the intelligibility of auditory (speech) input (Obleser and Weisz, 2012; Becker et al., 2013). Alpha power suppression was attenuated when auditory stimuli were degraded, that is, when comprehension was more effortful and required higher demands on attention (Obleser et al., 2012), as has been suggested for effortful listening situations before (e.g., Shinn-Cunningham and Best, 2008; Wild et al., 2012).

With respect to our data, we propose that alpha power increases gated the neural processing of acoustic information (duration vs. spectral peak) that differed in task-relevance between conditions: The introduction of spectral degradation in the second half of our experiment changed the relative informativeness or task-relevance of the spectral and duration cues, with spectral peak becoming less informative than stimulus duration. It is thus possible that enhanced alpha under degradation indexed the inhibition of spectral information processing.

Historically, however, enhanced alpha power has first been interpreted as reflecting the degree to which cortical areas are in an "idling" state (Adrian and Matthews, 1934; Niedermeyer and Silva, 2005). Consequently, reduction or suppression of alpha power was taken to index a departure from the idling mode toward a more attentive state. While this interpretation might be applicable for the general suppression of alpha power (vs. baseline) for nondegraded and degraded conditions, it cannot explain the differences in alpha power between conditions. That is, overall performance in our experiment (and thus presumably attentional effort) was comparable between the nondegraded and degraded conditions, while alpha power increased in the latter condition. Thus, this increase in alpha power is unlikely to reflect a more pronounced idling state.

Table 3 | Significant clusters obtained from whole-brain analyses ($p < 0.005$, extent threshold = 15) for the parametric modulators alpha and cue index, together with modulation differences between conditions.

Contrast	Area	Coordinates	*Z*	Extent (voxels)
Alpha power by BOLD (degraded)	r. oIFG/BA47	45, 29, −8	3.37	49
	r. IFG/BA45	54, 26, 10	3.25	16
	r. pSTG/PT	51, −43, 10	3.14	31
	r. AG/BA39	36, −67, 43	3.04	16
Alpha power by BOLD (nondegraded)	–	–	*n.s.*	
Alpha power degraded > nondegraded	r. oIFG/BA47	45, 29, −11	3.32	18
	r. pSTG/PT	54, −43, 13	3	15
	l. pSTG/PT	−54, −49, 13	2.94	22
Cue index by BOLD (degraded)	r. MFG	39, 47, 4	4.46	58
Cue index by BOLD (nondegraded)	–	–	*n.s.*	
Cue index degraded > nondegraded	r. DLPFC	42, 11, 28	3.74	49
	l. pSTG/PT	−54, −40, 7	3.7	21
	r. IPL	42, −40, 40	3.53	93
	l. MTG	−45, −55, 4	3.28	22

Abbreviations are explained in the text. Coordinates are given in MNI-space.

Finally, it has been recently proposed that alpha power enhancement can also be indicative of active processing states (Palva and Palva, 2011). According to the "active processing hypothesis," enhanced alpha power underlies the coordination of neural processing in task-relevant cortical structures, particularly for higher-order attentional and executive functions. Since the participants in our experiment seemed to be reluctant to refrain from spectral cue utilization under degradation, enhanced alpha power may also relate to "listening" harder for spectral cues, i.e., to an active process of utilizing spectral cues despite their being less informative. Both the "functional inhibition" and "active processing" hypotheses can be applied to the cortical regions in which alpha power positively correlated with BOLD.

SPECTRAL DEGRADATION AND THE PLANUM TEMPORALE

In the degraded condition of our experiment, we observed positive correlations of alpha power with BOLD activations in posterior STG and PT. The posterior STG and the PT have previously been suggested to subserve the processing of spectral information, and in particular, pitch and pitch changes (Zatorre et al., 1994; Zatorre and Belin, 2001; Schönwiesner et al., 2005; Hall and Plack, 2009; Alho et al., 2014). In particular, Hall and Plack (2009) provided evidence that apart from lateral Heschl's gyrus (Schneider et al., 2005; Warren et al., 2005), the (right) PT supports pitch processing to a substantial degree. Importantly, Hall and Plack (2009) used stimuli that bore close resemblance to our degraded sound stimuli such that participants may have perceived and processed pitch differences between our sound categories. Altogether, the involvement of pSTG and PT in our experiment is likely to reflect spectral processing. The positive correlation of alpha power and BOLD activation in this "hub"-like structure for auditory categorization (Griffiths and Warren, 2002) can shed further light onto the relative weighting of spectral vs. duration cues under degradation.

Previous studies using simultaneous EEG-fMRI recordings have observed positive and negative correlations of alpha power with BOLD (Laufs et al., 2003; Gonçalves et al., 2006; de Munck et al., 2007; Goldman et al., 2009; Scheeringa et al., 2009, 2011; Michels et al., 2010; Liu et al., 2012). The interpretation of negative correlations of alpha power with BOLD activations follows the functional inhibition hypothesis (Foxe et al., 1998; Klimesch et al., 2007; Foxe and Snyder, 2011; Weisz et al., 2011, 2013; Klimesch, 2012; Obleser and Weisz, 2012; Obleser et al., 2012). That is, regions where activations increase with decreasing alpha power have been suggested to be relevant for attending to informative stimulus features, while regions where alpha power is positively correlated with BOLD haven been suggested to support the suppression of non-informative (task-irrelevant) stimulus features. Positive correlations of alpha power with BOLD can also be interpreted within the "active processing hypothesis" (Palva and Palva, 2011). This hypothesis relates enhanced alpha power to stronger neural coordination in cortical areas processing task-relevant information, particularly for higher-order attentional and executive functions.

Here, we observed that the posterior STG and the PT showed increased activation for degraded vs. nondegraded stimuli, and that STG and PT activations positively correlated with alpha power. This can either be interpreted with the "functional inhibition hypothesis" or the "active processing hypothesis:"

According to the "functional inhibition hypothesis," the positive correlation of alpha power with BOLD activation in (right) PT may reflect the relative inhibition of spectral information in this brain area. In detail, introduction of spectral degradation affected the informativeness of spectral peak for categorization, and corresponded to a change in cue utilization. That is, spectral peak became relatively task-irrelevant, and may have been inhibited in pSTG and PT.

According to the "active processing hypothesis," the positive correlation of alpha power and BOLD activation in pSTG and PT (particularly under degradation) may reflect the enhanced need for neural coordination in order to maintain spectral cue utilization. Overall, cue indices remained negative even after spectral information was degraded, that is, participants still relied on their initial spectral categorization strategy. For maintenance of the spectral strategy, participants might have drawn on (right) posterior STG and PT resources. Thus, the positive correlation of alpha power and BOLD in these cortical regions may index the need to listen "harder" to degraded stimulus cues that once were informative.

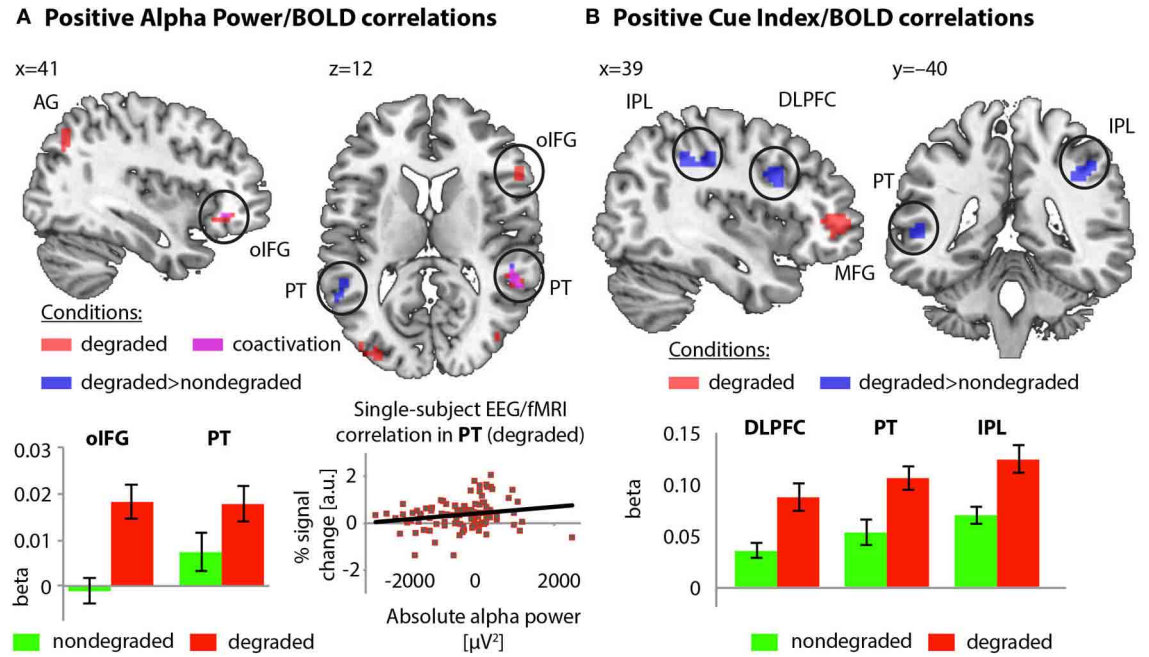


FIGURE 4 | (A) Positive correlation of alpha power with BOLD activity in the degraded condition (red), and correlation differences between conditions (blue; co-activation: magenta). Betas extracted from orbital inferior frontal gyrus and PT ROI visualize correlation differences between conditions. Data taken from a representative participant illustrate the positive single-subject alpha power/BOLD correlation in the degraded condition. **(B)** Positive correlations of cue index and BOLD in the degraded condition (red) and correlation differences between conditions (blue). Betas extracted from IFG, dorso-lateral prefrontal cortex and IPL visualize the correlation differences between conditions.

Finally, the "active processing hypothesis" seems to receive further support from the positive alpha power/BOLD correlations in frontal (IFG) areas. Note that Palva and Palva (2011) suggest that inhibition at lower sensory levels might be achieved by higher-level frontal functions, such that a positive alpha power/BOLD correlation in IFG may indicate that lesser reliance on spectral than on duration cues under degradation is mediated by activity in frontal regions. This may also relate to the observation that alpha power and behavioral cue utilization indices correlated only at trend-level with each other, suggesting that alpha power changes are more likely reflecting indirect, modulatory signatures of "functional inhibition" (after a stimulus while preparing a response, see also Obleser and Weisz, 2012; Wilsch et al., 2014). These signatures are dissociable from and follow in time early auditory signatures, accounting for the latency of the alpha power effect centered at around 500 ms post stimulus onset.

A ROLE OF THE RIGHT IPL IN AUDITORY ATTENTION

The behavioral tendency of disregarding spectral cues in the degraded condition of our experiment was accompanied by increased activation in anterior prefrontal cortex, and, compared to the nondegraded condition, in right IPL. In the degraded condition, right IPL showed a stronger correlation of cue index with BOLD activation than in the nondegraded condition (**Figure 4B**). As part of the fronto-parietal executive network (Posner and Dehaene, 1994; Corbetta et al., 2000), the IPL has repeatedly been found to subserve selective attention (Shaywitz et al., 2001; Behrmann et al., 2004; Salmi et al., 2009) and attentional control (Hill and Miller, 2010). Its activation was commonly observed in situations that require flexible changes in attention during the processing of informative stimulus features or task-relevant information (Geng and Mangun, 2009; Schultz and Lennert, 2009; Gillebert et al., 2012). In line with studies supporting the IPL's role in selectively attending to the most informative stimulus feature (Jacquemot et al., 2003; Gaab et al., 2006; Husain et al., 2006; Kiefer et al., 2008; Obleser et al., 2012), changes in IPL activation might support the change in cue utilization that was necessary for successful categorization (see Henry et al., 2013 for attention to temporal features). Note however that, behaviorally, participants tried to maintain their initial strategy and overall differed only marginally in cue utilization. Therefore, this interpretation must be considered carefully and substantiated by future research.

SUMMARY

In this multi-modal imaging study, we have shown that acoustic cue utilization during auditory categorization is flexible, even though listeners seem resilient to abandon initial categorization strategies. Brain areas processing the specific acoustic information—spectral peak vs. duration—supported the change in cue preference together with areas in the fronto-parietal attention network. Our data complement previous speech-related observations of alpha power increases in adverse and effortful listening situations (Obleser and Weisz, 2012; Obleser et al., 2012; Wilsch et al., 2014). We suggest that increased alpha power under degradation mediates the relative weighting of acoustic stimulus

features. Both the "functional inhibition" and the "active processing" hypotheses can account for these findings. Importantly, the combination of behavioral, electrophysiological, and hemodynamic measures is an indispensable methodology for further investigations in auditory cognition.

ACKNOWLEDGMENTS

Mathias Scharinger, Björn Herrmann, and Jonas Obleser are funded by the Max Planck Society. This research was supported by a Max Planck Research group grant to Jonas Obleser. We wish to express our special thanks to Dunja Kunke and Ina Koch for helping us with EEG preparations, to Sylvie Neubert for her help in participant recruitment and testing, and to Molly J. Henry and Thomas Gunter for helpful discussions and support.

REFERENCES

Adrian, E. D., and Matthews, B. H. C. (1934). The interpretation of potential waves in the cortex. *J. Physiol.* 81, 440–471.

Alho, K., Rinne, T., Herron, T. J., and Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: a meta-analysis of fMRI studies. *Hear. Res.* 307, 29–41. doi: 10.1016/j.heares.2013.08.001

Ashburner, J., and Friston, K. J. (2004). "Computational neuroanatomy," in *Human Brain Function*, eds R. S. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, C. Price, and S. Zeki (Amsterdam: Academic Press), 655–672.

Ashburner, J., and Good, C. D. (2003). "Spatial registration of images," in *Qualitative MRI of the Brain: Measuring Changes Caused by Disease*, ed P. Tofts (Chichester: John Wiley and Sons), 503–531. doi: 10.1002/0470869526.ch15

Becker, R., Pefkou, M., Michel, C. M., and Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Front. Syst. Neurosci.* 7:121. doi: 10.3389/fnsys.2013.00121

Behrmann, M., Geng, J. J., and Shomstein, S. (2004). Parietal cortex and attention. *Curr. Opin. Neurobiol.* 14, 212–217. doi: 10.1016/j.conb.2004.03.012

Bermudez, P., Lerch, J. P., Evans, A. C., and Zatorre, R. J. (2009). Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry. *Cereb. Cortex* 19, 1583–1596. doi: 10.1093/cercor/bhn196

Bertrand, O., and Pantev, C. (1994). "Stimulus frequency dependence of the transient oscillatory auditory evoked response (40 Hz) studied by electric and magnetic recordings in humans," in *Oscillatory Event-Related Brain Dynamics*, eds C. Pantev, T. Elbert, and B. Lütkenhöner (New York, NY: Plenum Press), 231–242. doi: 10.1007/978-1-4899-1307-4_17

Brett, M., Anton, J.-L., Valabregue, R., and Poline, J. B. (2002). "Region of interest analysis using an SPM toolbox," in *Paper Presented at the 8th International Conference on Functional Mapping of the Human Brain*, Sendai.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., and Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nat. Neurosci.* 3, 292–297. doi: 10.1038/73009

Debener, S., Strobel, A., Sorger, B., Peters, J., Kranczioch, C., Engel, A. K., et al. (2007). Improved quality of auditory event-related potentials recorded simultaneously with 3-T fMRI: removal of the ballistocardiogram artefact. *Neuroimage* 34, 587–597. doi: 10.1016/j.neuroimage.2006.09.031

Debener, S., Thorne, J., Schneider, T. R., and Viola, F. C. (2010). "Using ICA for the analysis of multi-channel EEG data," in *Simultaneous EEG and fMRI: Recording, Analysis, and Application*, eds M. Ullsperger and S. Debener (Oxford: Oxford University Press), 121–134. doi: 10.1093/acprof:oso/9780195372731.003.0008

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

de Munck, J. C., Gonçalves, S. I., Huijboom, L., Kuijer, J. P. A., Pouwels, P. J. W., Heethaar, R. M., et al. (2007). The hemodynamic response of the alpha rhythm: an EEG/fMRI study. *Neuroimage* 35, 1142–1151. doi: 10.1016/j.neuroimage.2007.01.022

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* 95, 1053–1064. doi: 10.1121/1.408467

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia* 50, 2154–2164. doi: 10.1016/j.neuropsychologia.2012.05.013

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *J. Neurosci.* 33, 10688–10697. doi: 10.1523/JNEUROSCI.4596-12.2013

Foxe, J. J., Simpson, G. V., and Ahlfors, S. P. (1998). Parieto-occipital approximately 10 Hz activity reflects anticipatory state of visual attention mechanisms. *Neuroreport* 9, 3929–3933. doi: 10.1097/00001756-199812010-00030

Foxe, J. J., and Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Front. Percept. Sci.* 2:154. doi: 10.3389/fpsyg.2011.00154

Friston, K. J. (2004). "Experimental design and statistical parametric mapping," in *Human Brain Function*, eds R. S. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, C. Price, and S. Zeki (Amsterdam: Academic Press), 599–632.

Gaab, N., Gaser, C., and Schlaug, G. (2006). Improvement-related functional plasticity following pitch memory training. *Neuroimage* 31, 255–263. doi: 10.1016/j.neuroimage.2005.11.046

Geng, J. J., and Mangun, G. R. (2009). Anterior intraparietal sulcus is sensitive to bottom-up attention driven by stimulus salience. *J. Cogn. Neurosci.* 21, 1584–1601. doi: 10.1162/jocn.2009.21103

Gillebert, C. R., Dyrholm, M., Vangkilde, S., Kyllingsbæk, S., Peeters, R., and Vandenberghe, R. (2012). Attentional priorities and access to short-term memory: parietal interactions. *Neuroimage* 62, 1551–1562. doi: 10.1016/j.neuroimage.2012.05.038

Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-T

Goldman, R. I., Stern, J. M., Engel, J. Jr., and Cohen, M. S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *Neuroreport* 13, 2487–2492. doi: 10.1097/00001756-200212200-00022

Goldman, R. I., Wei, C.-Y., Philiastides, M. G., Gerson, A. D., Friedman, D., Brown, T. R., et al. (2009). Single-trial discrimination for integrating simultaneous EEG and fMRI: identifying cortical areas contributing to trial-to-trial variability in the auditory oddball task. *Neuroimage* 47, 136–147. doi: 10.1016/j.neuroimage.2009.03.062

Gonçalves, S. I., de Munck, J. C., Pouwels, P. J. W., Schoonhoven, R., Kuijer, J. P. A., Maurits, N. M., et al. (2006). Correlating the alpha rhythm to BOLD using simultaneous EEG/fMRI: inter-subject variability. *Neuroimage* 30, 203–213. doi: 10.1016/j.neuroimage.2005.09.062

Goudbeek, M., Swingley, D., and Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1913–1933. doi: 10.1037/a0015781

Griffiths, T. D., and Warren, J. D. (2002). The planum temporale as a computational hub. *Trends Neurosci.* 25, 348–353. doi: 10.1016/S0166-2236(02)02191-4

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., and Tourville, J. A. (2004). Representation of sound categories in auditory cortical maps. *J. Speech Lang. Hear. Res.* 47, 46–57. doi: 10.1044/1092-4388(2004/005)

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). "Sparse temporal sampling" in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223.

Hall, D. A., Johnsrude, I. S., Haggard, M. P., Palmer, A. R., Akeroyd, M. A., and Summerfield, A. Q. (2002). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 12, 140–149. doi: 10.1093/cercor/12.2.140

Hall, D. A., and Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cereb. Cortex* 19, 576–585. doi: 10.1093/cercor/bhn108

Henry, M. J., Herrmann, B., and Obleser, J. (2013). Selective attention to temporal features on nested time scales. *Cereb. Cortex*. doi: 10.1093/cercor/bht240. [Epub ahead of print].

Herrmann, C. S., and Debener, S. (2008). Simultaneous recording of EEG and BOLD responses: a historical perspective. *Int. J. Psychophysiol.* 67, 161–168. doi: 10.1016/j.ijpsycho.2007.06.006

Hill, K. T., and Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cereb. Cortex* 20, 583–590. doi: 10.1093/cercor/bhp124

Holt, L. L., and Lotto, A. J. (2006). Cue weighting in auditory categorization: implications for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059–3071. doi: 10.1121/1.2188377

Husain, F. T., Fromm, S. J., Pursley, R. H., Hosey, L. A., Braun, A. R., and Horwitz, B. (2006). Neural bases of categorization of simple speech and nonspeech sounds. *Hum. Brain Mapp.* 27, 636–651. doi: 10.1002/hbm.20207

Huster, R. J., Debener, S., Eichele, T., and Herrmann, C. S. (2012). Methods for simultaneous EEG-fMRI: an introductory review. *J. Neurosci.* 32, 6053–6060. doi: 10.1523/JNEUROSCI.0447-12.2012

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., and Turner, R. (2002). Image distortion correction in fMRI: a quantitative evaluation. *Neuroimage* 16, 217–240. doi: 10.1006/nimg.2001.1054

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., and Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *J. Neurosci.* 23, 9541–9546.

Jann, K., Dierks, T., Boesch, C., Kottlow, M., Strik, W., and Koenig, T. (2009). BOLD correlates of EEG alpha phase-locking and the fMRI default mode network. *Neuroimage* 45, 903–916. doi: 10.1016/j.neuroimage.2009.01.001

Jensen, O., and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. Hum. Neurosci.* 4:186. doi: 10.3389/fnhum.2010.00186

Jezzard, P., and Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magn. Reson. Med.* 34, 65–73. doi: 10.1002/mrm.1910340111

Jokisch, D., and Jensen, O. (2007). Modulation of gamma and alpha activity during a working memory task engaging the dorsal or ventral stream. *J. Neurosci.* 27, 3244–3251. doi: 10.1523/JNEUROSCI.5399-06.2007

Kerlin, J. R., Shahin, A. J., and Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party." *J. Neurosci.* 30, 620–628. doi: 10.1523/JNEUROSCI.3631-09.2010

Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., and Hoenig, K. (2008). The sound of concepts: four markers for a link between auditory and conceptual brain systems. *J. Neurosci.* 28, 12224–12230. doi: 10.1523/JNEUROSCI.3579-08.2008

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends Cogn. Sci.* 16, 606–617. doi: 10.1016/j.tics.2012.10.007

Klimesch, W., Sauseng, P., and Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition-timing hypothesis. *Brain Res. Rev.* 53, 63–88. doi: 10.1016/j.brainresrev.2006.06.003

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., et al. (2003). EEG-correlated fMRI of human alpha activity. *Neuroimage* 19, 1463–1476. doi: 10.1016/S1053-8119(03)00286-6

Liu, Z., de Zwart, J. A., Yao, B., van Gelderen, P., Kuo, L.-W., and Duyn, J. H. (2012). Finding thalamic BOLD correlates to posterior alpha EEG. *Neuroimage* 63, 1060–1069. doi: 10.1016/j.neuroimage.2012.08.025

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Mahwah, NJ: Erlbaum.

Michels, L., Bucher, K., Lüchinger, R., Klaver, P., Martin, E., Jeanmonod, D., et al. (2010). Simultaneous EEG-fMRI during a working memory task: modulations in low and high frequency bands. *PLoS ONE* 5:e10298. doi: 10.1371/journal.pone.0010298

Moosmann, M., Ritter, P., Krastel, I., Brink, A., Thees, S., Blankenburg, F., et al. (2003). Correlates of alpha rhythm in functional magnetic resonance imaging and near infrared spectroscopy. *Neuroimage* 20, 145–158. doi: 10.1016/S1053-8119(03)00344-6

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Nahum, M., Nelken, I., and Ahissar, M. (2008). Low-level information and high-level perception: the case of speech in noise. *PLoS Biol.* 6:e216. doi: 10.1371/journal.pbio.0060126

Niedermeyer, E., and Silva, F. H. L. D. (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Philadelphia, PA: Lippincott Williams and Wilkins.

Obleser, J., and Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends Cogn. Sci.* 13, 14–19. doi: 10.1016/j.tics.2008.09.005

Obleser, J., and Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cereb. Cortex* 20, 633–640. doi: 10.1093/cercor/bhp128

Obleser, J., and Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cereb. Cortex* 22, 2466–2477. doi: 10.1093/cercor/bhr325

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., and Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *J. Neurosci.* 32, 12376–12383. doi: 10.1523/JNEUROSCI.4908-11.2012

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

Oostenveld, R., and Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clin. Neurophysiol.* 112, 713–719. doi: 10.1016/S1388-2457(00)00527-7

Palva, S., and Palva, J.M. (2011). Functional roles of alpha-band phase synchronization in local and large-scale cortical networks. *Front. Psychol.* 2:204. doi: 10.3389/fpsyg.2011.00204

Peelle, J. E., Eason, R. J., Schmitter, S., Schwarzbauer, C., and Davis, M. H. (2010). Evaluating an acoustically quiet EPI sequence for use in fMRI studies of speech and auditory processing. *Neuroimage* 52, 1410–1419. doi: 10.1016/j.neuroimage.2010.05.015

Poissant, S. F., Whitmal, N. A. 3rd., and Freyman, R. L. (2006). Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *J. Acoust. Soc. Am.* 119, 1606–1615. doi: 10.1121/1.2168428

Posner, M. I., and Dehaene, S. (1994). Attentional networks. *Trends Neurosci.* 17, 75–79. doi: 10.1016/0166-2236(94)90078-7

Rihs, T. A., Michel, C. M., and Thut, G. (2007). Mechanisms of selective inhibition in visual spatial attention are indexed by alpha-band EEG synchronization. *Eur. J. Neurosci.* 25, 603–610. doi: 10.1111/j.1460-9568.2007.05278.x

Rinne, T., Stecker, G. C., Kang, X., Yund, E. W., Herron, T. J., and Woods, D. L. (2007). Attention modulates sound processing in human auditory cortex but not the inferior colliculus. *Neuroreport* 18, 1311–1314. doi: 10.1097/WNR.0b013e32826fb3bb

Ritter, P., and Villringer, A. (2006). Simultaneous EEG-fMRI. *Neurosci. Biobehav. Rev.* 30, 823–838. doi: 10.1016/j.neubiorev.2006.06.008

Ronnberg, 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, S99–S105. doi: 10.1080/14992020802301167

Rosen, S., Faulkner, A., and Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: implications for cochlear implants. *J. Acoust. Soc. Am.* 106, 3629–3636. doi: 10.1121/1.428215

Sadaghiani, S., Scheeringa, R., Lehongre, K., Morillon, B., Giraud, A.-L., D'Esposito, M., et al. (2012). Alpha-band phase synchrony is related to activity in the fronto-parietal adaptive control network. *J. Neurosci.* 32, 14305–14310. doi: 10.1523/JNEUROSCI.1358-12.2012

Sadaghiani, S., Scheeringa, R., Lehongre, K., Morillon, B., Giraud, A.-L., and Kleinschmidt, A. (2010). Intrinsic connectivity networks, alpha oscillations, and tonic alertness: a simultaneous electroencephalography/functional magnetic resonance imaging study. *J. Neurosci.* 30, 10243–10250. doi: 10.1523/JNEUROSCI.1004-10.2010

Salmi, J., Rinne, T., Koistinen, S., Salonen, O., and Alho, K. (2009). Brain networks of bottom-up triggered and top-down controlled shifting of auditory attention. *Brain Res.* 1286, 155–164. doi: 10.1016/j.brainres.2009.06.083

Scharinger, M., Henry, M. J., Erb, J., Meyer, L., and Obleser, J. (2014). Thalamic and parietal brain morphology predicts auditory category learning. *Neuropsychologia* 53, 75–83. doi: 10.1016/j.neuropsychologia.2013.09.012

Scharinger, M., Henry, M. J., and Obleser, J. (2013). Prior experience with negative spectral correlations promotes information integration during auditory category learning. *Mem. Cogn.* 41, 752–768. doi: 10.3758/s13421-013-0294-9

Scheeringa, R., Fries, P., Petersson, K.-M., Oostenveld, R., Grothe, I., Norris, D. G., et al. (2011). Neuronal dynamics underlying high- and low-frequency EEG oscillations contribute independently to the human BOLD signal. *Neuron* 69, 572–583. doi: 10.1016/j.neuron.2010.11.044

Scheeringa, R., Petersson, K. M., Kleinschmidt, A., Jensen, O., and Bastiaansen, M. C. M. (2012). EEG alpha power modulation of fMRI resting-state connectivity. *Brain Connect.* 2, 254–264. doi: 10.1089/brain.2012.0088

Scheeringa, R., Petersson, K. M., Oostenveld, R., Norris, D. G., Hagoort, P., and Bastiaansen, M. C. M. (2009). Trial-by-trial coupling between EEG and BOLD identifies networks related to alpha and theta EEG power increases during working memory maintenance. *Neuroimage* 44, 1224–1238. doi: 10.1016/j.neuroimage.2008.08.041

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., et al. (2005). Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nat. Neurosci.* 8, 1241–1247. doi: 10.1038/nn1530

Schönwiesner, M., Rübsamen, R., and von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *Eur. J. Neurosci.* 22, 1521–1528. doi: 10.1111/j.1460-9568.2005.04315.x

Schultz, J., and Lennert, T. (2009). BOLD signal in intraparietal sulcus covaries with magnitude of implicitly driven attention shifts. *Neuroimage* 45, 1314–1328. doi: 10.1016/j.neuroimage.2009.01.012

Scott, S. K., Rosen, S., Lang, H., and Wise, R. J. S. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech–a positron emission tomography study. *J. Acoust. Soc. Am.* 120, 1075–1083. doi: 10.1121/1.2216725

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Sharda, M., and Singh, N. C. (2012). Auditory perception of natural sound categories - An fMRI study. *Neuroscience* 214, 49–58. doi: 10.1016/j.neuroscience.2012.03.053

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Fulbright, R. K., Skudlarski, P., Mencl, W. E., et al. (2001). The functional neural architecture of components of attention in language-processing tasks. *Neuroimage* 13, 601–612. doi: 10.1006/nimg.2000.0726

Shinn-Cunningham, B. G., and Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplif.* 12, 283–299. doi: 10.1177/1084713808325306

Slotnick, S. D., Moo, L. R., Segal, J. B., and Hart, J. Jr. (2003). Distinct prefrontal cortex activity associated with item memory and source memory for visual shapes. *Brain Res. Cogn. Brain Res.* 17, 75–82. doi: 10.1016/S0926-6410(03)00082-X

Smits, R., Sereno, J., and Jongman, A. (2006). Categorization of sounds. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 733–754. doi: 10.1037/0096-1523.32.3.733

Thut, G., Nietzel, A., Brandt, S. A., and Pascual-Leone, A. (2006). Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. *J. Neurosci.* 26, 9494–9502. doi: 10.1523/JNEUROSCI.0875-06.2006

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Warren, J. D., Jennings, A. R., and Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *Neuroimage* 24, 1052–1057. doi: 10.1016/j.neuroimage.2004.10.031

Weissman, D. H., Warner, L. M., and Woldorff, M. G. (2009). Momentary reductions of attention permit greater processing of irrelevant stimuli. *Neuroimage* 48, 609–615. doi: 10.1016/j.neuroimage.2009.06.081

Weisz, N., Hartmann, T., Müller, N., Lorenz, I., and Obleser, J. (2011). Alpha rhythms in audition: cognitive and clinical perspectives. *Front. Psychol.* 2:73. doi: 10.3389/fpsyg.2011.00073

Weisz, N., Müller, N., Jatzev, S., and Bertrand, O. (2013). Oscillatory alpha modulations in right auditory regions reflect the validity of acoustic cues in an auditory spatial attention task. *Cereb. Cortex*. doi: 10.1093/cercor/bht113. [Epub ahead of print].

Westbury, C. F., Zatorre, R. J., and Evans, A. C. (1999). Quantifying variability in the planum temporale: a probability map. *Cereb. Cortex* 9, 392–405. doi: 10.1093/cercor/9.4.392

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

Wilsch, A., Henry, M. J., Herrmann, B., Maess, B., and Obleser, J. (2014). Alpha oscillatory dynamics index temporal expectation benefits in working memory. *Cereb. Cortex*. doi: 10.1093/cercor/bhu1004. [Epub ahead of print].

Yantis, S. (1993). Stimulus-driven attentional capture and attentional control settings. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 676–681. doi: 10.1037/0096-1523.19.3.676

Yantis, S. (2008). The neural basis of selective attention: cortical sources and targets of attentional modulation. *Curr. Dir. Psychol. Sci.* 17, 86–90. doi: 10.1111/j.1467-8721.2008.00554.x

Zatorre, R. J., and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cereb. Cortex* 11, 946–953. doi: 10.1093/cercor/11.10.946

Zatorre, R. J., Evans, A. C., and Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *J. Neurosci.* 14, 1908–1919.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

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

Received: 24 February 2014; accepted: 17 May 2014; published online: 04 June 2014.
*Citation: Scharinger M, Herrmann B, Nierhaus T and Obleser J (2014) Simultaneous EEG-fMRI brain signatures of auditory cue utilization. Front. Neurosci. **8**:137. doi: 10.3389/fnins.2014.00137*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

REVIEW ARTICLE
published: 01 August 2014
doi: 10.3389/fnins.2014.00230

How may the basal ganglia contribute to auditory categorization and speech perception?

***Sung-Joo Lim*[1,2]*, *Julie A. Fiez*[2,3,4] *and Lori L. Holt*[1,2,3]**

[1] *Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA*
[2] *Department of Neuroscience, Center for the Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, PA, USA*
[3] *Department of Neuroscience, Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA*
[4] *Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA*

Edited by:
Einat Liebenthal, Medical College of Wisconsin, USA

Reviewed by:
Carol Seger, Colorado State University, USA
Ingo Hertrich, University of Tuebingen, Germany

****Correspondence:***
Sung-Joo Lim, Auditory Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, 04103 Leipzig, Germany
e-mail: sungjoo@cbs.mpg.de

Listeners must accomplish two complementary perceptual feats in extracting a message from speech. They must discriminate linguistically-relevant acoustic variability and generalize across irrelevant variability. Said another way, they must *categorize* speech. Since the mapping of acoustic variability is language-specific, these categories must be learned from experience. Thus, understanding how, in general, the auditory system acquires and represents categories can inform us about the toolbox of mechanisms available to speech perception. This perspective invites consideration of findings from cognitive neuroscience literatures outside of the speech domain as a means of constraining models of speech perception. Although neurobiological models of speech perception have mainly focused on cerebral cortex, research outside the speech domain is consistent with the possibility of significant subcortical contributions in category learning. Here, we review the functional role of one such structure, the basal ganglia. We examine research from animal electrophysiology, human neuroimaging, and behavior to consider characteristics of basal ganglia processing that may be advantageous for speech category learning. We also present emerging evidence for a direct role for basal ganglia in learning auditory categories in a complex, naturalistic task intended to model the incidental manner in which speech categories are acquired. To conclude, we highlight new research questions that arise in incorporating the broader neuroscience research literature in modeling speech perception, and suggest how understanding contributions of the basal ganglia can inform attempts to optimize training protocols for learning non-native speech categories in adulthood.

Keywords: speech category learning, perceptual learning, basal ganglia, speech perception, categorization, plasticity

INTRODUCTION

Speech is a highly variable signal. A central challenge for listeners is discovering how this variability maps to language. A change in pitch may be a linguistically irrelevant deviation arising from emotion, or a telling acoustic cue to whether the sound signaled *beach* or *peach*. This is an example of *categorization*, in that potentially discriminable sounds come to be treated as functionally equivalent classes defined by relevant features (see Holt and Lotto, 2010, for a review). Because this perceptual mapping of sounds is specific to linguistic categories (e.g., consonant and vowel phonemes), one must learn speech categories through experience with the native language. Infants begin to learn native-language speech categories within their first year; exposure to native speech input warps speech perception, enhancing discrimination across native speech categories but diminishing within-category discrimination (Kuhl et al., 1992, 2006), and discrimination of non-native categories not present in the native language (Werker and Tees, 1984). By adulthood, one becomes "neurally committed" to native-language-specific speech categories (see Kuhl, 2004, for a review), which in turn can lead to profound difficulty in learning non-native speech categories as an adult (Best, 1995; Flege, 1995). This pattern indicates that experience with the native language plays a crucial role in shaping how we perceive speech.

However, relatively less is known about *how* speech categories are acquired through experience. One main challenge to our understanding is gaining experimental control over participants' history of linguistic experience. Adult listeners' perception has already been tuned by long-term native speech experience, the extent of which cannot be fully measured by the experimenter. Likewise, it is impossible to determine even young infants' speech experience. Exposure to native-language speech is substantial in the early postnatal months and speech experience begins even prenatally (Mehler et al., 1988; Moon et al., 1993). This lack of experimental control imposes critical limitations on understanding of the role of language experience on speech category acquisition, and impedes development of a mechanistic framework of how speech categories are learned.

A small, but growing, literature has been motivated by the premise that modeling the challenges of speech category learning using nonspeech sounds can reveal principles of general auditory

category learning. Understanding these principles reveals characteristics of auditory learning available to support speech category learning. For instance, by using novel nonspeech sound categories, Holt and Lotto (2006) demonstrated that distributional characteristics of sound category input influence listeners' perceptual weighting of multiple acoustic cues for categorization. This finding led Lim and Holt (2011) to test whether increasing variability along a cue that is inefficient in a second language may lead second language learners to rely upon it less in subsequent speech categorization. They found that in Japanese adults learning English, increasing the distributional variance along the native Japanese listeners' preferred (but non-diagnostic for English) acoustic cue led the listeners to rely on this cue less in subsequent English speech categorization. This example demonstrates that learning about general auditory categorization processes can inform our approaches to understanding speech perception and learning.

This general perspective on speech perception invites consideration of findings from the cognitive neuroscience literature outside of the domain of speech and auditory processing. Parallel lines of general learning research suggest that there are multiple learning systems and corresponding neural structures, with an emphasis on the significant contributions of subcortical structures in learning (e.g., Doya, 1999, 2000; Ashby and O'Brien, 2005; Seger and Miller, 2010). Understanding the involvement of subcortical learning systems is especially important to developing full neurobiological models of speech categorization, because current neurobiological and theoretical models of speech processing have focused mainly on the cerebral cortex (McClelland and Elman, 1986; Hickok and Poeppel, 2004; but see Guenther, 1995; Guenther and Ghosh, 2003; Guediche et al., 2014).

In the present review, we focus on the potential of one such subcortical system—the basal ganglia—to play a role in speech categorization. The basal ganglia have been widely implicated in category learning outside the domain of speech processing. Basal ganglia-mediated category learning research, conducted mostly in the domain of visual categorization, has focused on learning mechanisms at the level of category decision-making (i.e., selecting appropriate motor responses associated with category membership). This contrasts to the general approach in speech categorization research, which has focused largely on learning-induced category representations occurring at the sensory level (e.g., Callan et al., 2003; Golestani and Zatorre, 2004; Liebenthal et al., 2005; Desai et al., 2008; Lee et al., 2012). It is important to note that these differing perspectives likely represent attention to different aspects of a larger system. Thus, they are potentially mutually informative, although as of yet they have not been integrated in the service of understanding categorization. Here, we aim to review these different lines of research from the perspective of how they can inform speech categorization.

We begin by reviewing the functional role of the basal ganglia. We examine research from animal electrophysiology, human neuroimaging, and human behavior to identify characteristics of basal ganglia processing that may be advantageous for speech category learning. We then consider the basal ganglia as a system that may play a role in auditory category learning. We focus on characteristics that can potentially contribute to learning of speech categories and training approaches to promote effective non-native speech category acquisition.

OVERVIEW OF THE BASAL GANGLIA AND REINFORCEMENT LEARNING

The basal ganglia are a collection of subcortical nuclei with a complex circuitry. The input nuclei of the basal ganglia consist of the caudate nucleus and putamen (together referred to as the dorsal striatum) and the nucleus accumbens (considered part of the ventral striatum). The dorsal and ventral striatum receive input from the cerebral cortex and send projections to the output nuclei of the basal ganglia, which include the globus pallidus and the substantia nigra pars reticulata (see **Figure 1**). The output signals from these nuclei ultimately project back to the cerebral cortex via the thalamus (see **Figure 2**). This basal ganglia-thalamo-cortical circuitry forms "closed loops," whereby cortical regions projecting to the basal ganglia receive recurrent feedback projections from the basal ganglia (Alexander et al., 1986) and also "open loops," whereby cortical regions projecting to the basal ganglia terminate in different cortical regions via the basal ganglia (Joel and Weiner, 1994). In addition to these structures, neurons in the substantia nigra pars compacta and ventral tegmental area play a crucial role in mediating basal ganglia's functions. Dopamingeric projections from these neurons modulate activity of the dorsal and ventral striatum, which ultimately modulate plasticity among the synapses within basal ganglia-thalamo-cortical loops (Reynolds and Wickens, 2002).

The traditional view holds that the basal ganglia are mostly involved in motor-related processing and learning. Basal ganglia circuitry was thought to mainly innervate the primary motor cortex (Kemp and Powell, 1971), which could account for the pronounced movement-related deficits commonly observed among patients with diseases that damage the basal ganglia (e.g., Parkinson's and Huntington's diseases). However, more recent findings have indicated that the basal ganglia nuclei are highly interconnected with widespread areas of the cerebral cortex

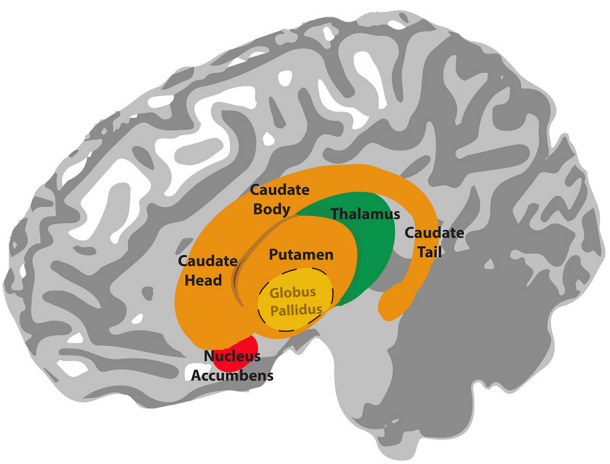


FIGURE 1 | Illustration of the anatomy of the basal ganglia. The globus pallidus lies inside the putamen. The thalamus is located underneath the basal ganglia, in the medial position of the brain.

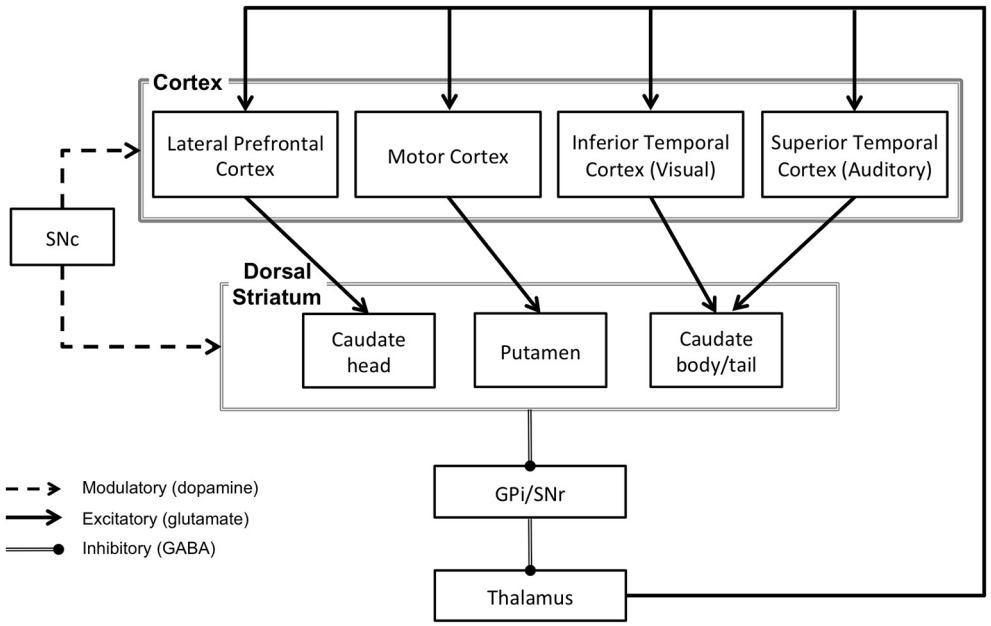


FIGURE 2 | The direct pathway circuitry of the basal ganglia via the dorsal striatum. SNc, substantia nigra pars compacta; SNr, substantia nigra pars reticulata; GPi, globus pallidus, internal portion.

(Alexander et al., 1986; Middleton and Strick, 2000). This view suggests that the basal ganglia not only influence motor-related processes, but also play an important role in non-motor cognitive functions and a wide range of learning challenges, including perceptual categorization (e.g., Ashby et al., 1998; Hochstenbach et al., 1998; see Lawrence et al., 1998; Saint-Cyr, 2003; Seger, 2008, for reviews).

The basal ganglia are crucially involved in learning appropriate behavioral actions to achieve goals in a given environment. This type of learning can be explained by a computational theory, reinforcement learning, whereby learning emerges as one builds and updates predictions about receiving future rewards. Learning occurs in minimizing the difference between predictions of reward and actual reward, referred as a reward prediction error (Sutton and Barto, 1998). In this way, an unexpected reward or punishment is an indicator that the value of an environmental stimulus (or the best response to it) was not accurately predicted. Therefore, errors in predictions lead to adjustments to predicted value and stimulus-action associations. Based on such predictions, behavior adjusts adaptively to maximize future rewards such that actions leading to rewards are reinforced (i.e., the likelihood of the specific actions increases), whereas incorrect behaviors leading to punishment (or no rewards) are modified. Through this process, reward drives learning of goal-directed actions thereby shaping behavior.

The basal ganglia have been implicated in reinforcement learning by means of the neuromodulatory activity of dopamine neurons located in the midbrain (Schultz et al., 1997; Schultz, 1999; Daw et al., 2005). The dopamine neurons that project to the dorsal striatum are located in the substantia nigra (the pars compacta sector), whereas those that project to the ventral striatum are located in the ventral tegmental area (Nauta et al., 1974; Simon et al., 1979; Swanson, 1982; Amalric and Koob, 1993; Haber and Fudge, 1997). Electrophysiological recording studies on primates by Shultz and colleagues (Schultz et al., 1993, 1997) indicate that dopamine neurons are sensitive to reward prediction. These studies have shown that in the initial phase of learning when rewards are not expected, dopamine neurons fire (i.e., release dopamine) at the onset of reward delivery, but over the course of learning these neurons begin to fire to cues that predict rewarding outcome. When an expected reward is omitted or fails to occur, dopamine levels are depressed (Schultz et al., 1997; Hollerman and Schultz, 1998; Schultz, 1998). A similar pattern of reward-related dopamine neuronal firing is reflected in the activity in the striatum (Hikosaka et al., 1989; Robbins and Everitt, 1992; Schultz et al., 1992, 1993; Tremblay et al., 1998; Schultz, 2000; Berns et al., 2001; McClure et al., 2003).

Computationally, the observed patterns of activity are consistent with the idea that dopamine neurons can signal reward prediction error, which can serve as a teaching signal to drive reinforcement learning. The presumed reward prediction error signals carried by dopamine neurons are thought to modulate the synaptic plasticity of cortico-striatal pathways (Reynolds and Wickens, 2002). Dopamine release can induce long-term potentiation, which effectively strengthens cortico-striatal synapses at the site of release (Wickens et al., 1996; Kerr and Wickens, 2001). This process may be significant in strengthening striatal pathways that encode contexts that predict reward and promote learning of goal-directed actions (i.e., stimulus-response-outcome associations). Therefore, dopamine may be regarded as a learning signal (e.g., Beninger, 1983; Wise and Rompre, 1989; Wickens, 1997; Schultz, 1998, 2002) that reinforces rewarding actions

by strengthening stimulus-action associations (Law of Effect, Thorndike, 1911) and mediating relevant cortico-striatal loops to accomplish learning (Houk and Wise, 1995). Conversely, in the case of punishment or omission of expected reward, a relative depression of dopamine levels would induce long-term depression, thus weakening the synapses (Wickens et al., 2003; Calabresi et al., 2007). It is of note that dopamine-mediated learning does not necessarily occur solely through reward prediction error signals processed via the striatum, since dopamine neurons also send direct projections to the cortex (Thierry et al., 1973; Hökfelt et al., 1974, 1977; Lindvall et al., 1974; see Foote and Morrison, 1987, for a review). Nevertheless, the dopaminergic signals through the striatum are likely to be a more robust learning signal, since dopamine neurons disproportionately project to the striatum (Szabo, 1979; Selemon and Goldman-Rakic, 1990; Hedreen and DeLong, 1991; Lynd-Balta and Haber, 1994).

The findings in non-human primates converge with evidence from human neuroimaging studies. Across various learning tasks, including learning non-native phonetic categories (Tricomi et al., 2006), it has been found that activity in the dorsal striatum is modulated according to the valence and the value of feedback that is contingent to one's response actions (i.e., goal-directed behavior) (Elliott et al., 1997, 2004; Koepp et al., 1998; Delgado et al., 2000, 2004; Haruno et al., 2004; O'Doherty et al., 2004; Tricomi et al., 2006). Yet, it is significant to note that rather than responding to response outcomes *per se*, the dorsal striatum exhibits greater activity when individuals perceive the outcomes as contingent on their actions and relevant to their goals (i.e., receiving reward) (Tricomi et al., 2004; Tricomi and Fiez, 2008). Surprisingly, the striatum can even show a reward-like response to negative feedback, if this feedback provides useful information for predicting future rewards (Tricomi and Fiez, 2012). This demonstrates that the striatum is sensitive to the subjective value of information for goal achievements (Tricomi and Fiez, 2008; Han et al., 2010). More generally, these findings suggest that reinforcement learning in humans involves the striatum and it extends into the cognitive domain, as learning can be influenced by high-level thought processes relating to motivation and goal-directed actions.

CONTRIBUTIONS OF THE BASAL GANGLIA TO NON-NATIVE SPEECH CATEGORY LEARNING

In this section, we consider the challenges involved in learning non-native speech categories and the relative ineffectiveness of passive exposure to non-native speech to improve categorization performance. Then, we review evidence for the effectiveness of directed category training, in which individuals receive goal-relevant feedback about the accuracy of their category judgments. We consider evidence that such training involves an anterior basal ganglia system that drives learning-related changes in non-native speech categorization. Finally, we examine the limitations of directed category training, and consider whether training that encourages the use of procedural learning mechanisms involving a posterior basal ganglia system may be more suited for the perceptual demands of speech category learning.

Adults find it notoriously difficult to learn some non-native speech categories even with extensive training or years of exposure to a foreign language (Gordon et al., 2001; Aoyama et al., 2004; Ingvalson et al., 2011). This difficulty is partly due to interference from expertise with native-language speech categories (Best, 1995; Flege, 1995) developed from long-term experience with their native language since infancy (Werker and Tees, 1984). The case of native Japanese adults' acquisition of English /r/-/l/ has been a prominent example of the difficulty acquiring some non-native speech categories (Goto, 1971; Miyawaki et al., 1975; Werker and Logan, 1985). Whereas English divides the perceptual space into two phonetic categories, /r/ and /l/ as in *rock* and *lock*, there is a single Japanese speech category within a similar perceptual space (Lotto et al., 2004). Having learned this single Japanese category, native Japanese adults have great difficulty distinguishing English /r/-/l/ due to the persistent reliance on the native Japanese perceptual space (Iverson et al., 2003). This difficulty presents important questions regarding the limits and challenges to perceptual plasticity in adulthood.

In attempts to understand adult second language speech category learning, different types of laboratory-controlled training tasks have been used. One common task is unsupervised listening, in which listeners are passively exposed to sound stimuli. Studies using this type of task have shown that listeners' perception is tuned according to the statistical regularity in the input; they become sensitive to the distributional regularities of speech syllables (Maye et al., 2002; Clayards et al., 2008; Goudbeek et al., 2008), correlations between acoustic features defining the units (Idemaru and Holt, 2011), and sequential relationships between syllabic units or tones (Saffran et al., 1996, 1999). However, this type of training fails to facilitate non-native speech category learning in adults. McClelland and colleagues (McClelland et al., 1999; McCandliss et al., 2002; Vallabha and McClelland, 2007) argue that English /r/ and /l/ exemplars are perceptually similar enough to the single Japanese category that hearing English /r/ and /l/ tends to simply activate and strengthen the Japanese category representation among native Japanese adults. They argue that this arises from Hebbian learning principles interacting with the perceptual organization brought about by Japanese language experience. Therefore, unsupervised learning of non-native speech categories may fail unless special steps are taken, such as artificially exaggerating the training stimuli so that they can be perceived as distinct category instances (McCandliss et al., 2002; Tricomi et al., 2006; Ingvalson et al., 2011).

The other dominant, perhaps more effective, training approach to achieve non-native speech category learning is to use directed training that requires overt categorization or identification responses and provides explicit trial-by-trial feedback about the correctness of the response. Directed categorization training has been commonly used to investigate non-native speech category learning (e.g., Logan et al., 1991; Lively et al., 1993, 1994; Bradlow et al., 1997; Wang et al., 1999; Iverson et al., 2005; Francis et al., 2008). Comparisons between passive exposure and directed training tasks have demonstrated an advantage for directed training in learning auditory and speech categories (McCandliss et al., 2002; McClelland et al., 2002; Goudbeek et al., 2008). Although previous training studies have focused on the impact of the acoustic characteristics of training stimuli on learning (Logan et al., 1991; Lively et al., 1993, 1994; Iverson et al., 2005), the learning

advantage observed for directed training over passive listening tasks indicates that the details of training are crucial.

Using fMRI, Tricomi et al. (2006) demonstrated that directed category training of non-native speech categories engages the basal ganglia (i.e., the striatum), as compared to a condition without performance feedback. The findings illustrated that the nature of the training task engaged different neural processes and learning systems. Performance feedback may potentially play a crucial role in informing the *functional distinctiveness* of non-native speech categories in traditional laboratory training tasks. Through corrective feedback that encourages distinct action associations (e.g., button presses) for the categories, one's actions are shaped to respond differently to these sound categories, thereby assigning distinct behavioral significance to the sounds.

It is notable that non-native speech category learning in adulthood occurs with directed categorization training, but learning gains are relatively modest even across multiple weeks of extensive training (e.g., Logan et al., 1991; Lively et al., 1993; Bradlow et al., 1997; Iverson et al., 2005). Given the literature reviewed above, which demonstrates that task and stimulus details can be influential in engaging different learning systems, there is the possibility that overt categorization tasks with explicit feedback may fail to tap into the most effective learning mechanisms for adult speech category learning.

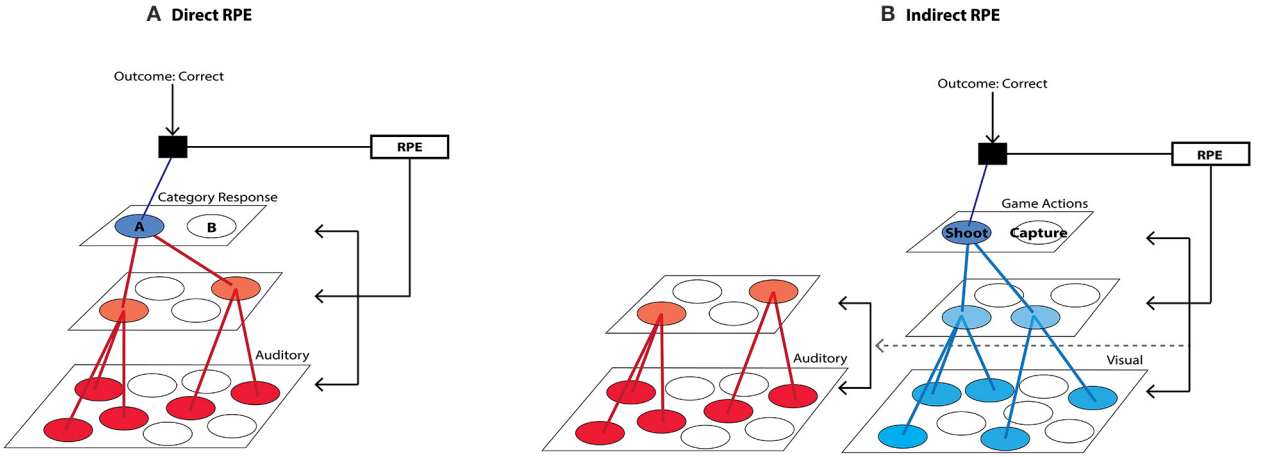
One of the main challenges of speech perception and categorization is to map highly variable sound exemplars distributed across multiple acoustic dimensions onto linguistically-relevant phonemic categories (see Holt and Lotto, 2010, for a review). Speech categories are inherently multidimensional such that no single acoustic cue or dimension is sufficient to define category membership. For example, Lisker (1986) has reported that there are as many as 16 acoustic cues, all of which can be used to distinguish voiced vs. voiceless consonants (e.g., /ba/ vs. /pa/). Therefore, listeners must integrate multiple acoustic cues for speech categorization (Liberman et al., 1967; Liberman, 1996). Furthermore, there is high variability in these acoustic cues originating from different speech contexts, speaker's characteristics, among other sources. Adding to this complexity, temporal transitions of these acoustic cues occur at a millisecond scale that requires rapid tracking of simultaneous acoustic dimensions. These characteristics of the speech signal make it difficult to acquire explicit knowledge about the crucial acoustic dimensions that define speech categories. Therefore, learning of speech categories essentially represents learning of procedural knowledge that cannot be explicitly verbalized.

Since speech perception and learning inherently require integration of multiple, highly varying acoustic dimensions, explicit attempts to discover and integrate acoustic cues that are diagnostic to speech category identity may be extremely difficult. Yet, it has been shown that directed categorization training is likely to engage explicit/directed attention to acoustic features (Logan et al., 1991), and to recruit a sector of the basal ganglia (the head of the caudate nucleus) implicated in executive control and the cognitive processing of feedback (Tricomi et al., 2006). Learners are aware of the relationship between the outcome and speech categories in directed categorization training. Thus, they may attempt to discover potential features that may be critical for categorization in a declarative manner, which might not be optimal for learning speech categories due to their complex, difficult-to-verbalize nature (see **Box 1A**).

Within the domain of visual categorization, Ashby and colleagues have suggested that learning verbal rules (i.e., declarative knowledge) vs. integration of dimensions (i.e., procedural knowledge) that define categories is achieved by distinct, competitive learning systems (Ashby et al., 1998; Ashby and Ell, 2001; Ashby and Maddox, 2005). Learning declarative knowledge about the category features that are verbalizable engages executive attention and working memory, mediated by the prefrontal cortex and the anterior portion of the dorsal striatum (i.e., the head of the caudate nucleus). In contrast, acquisition of novel visual categories that require integration of multiple stimulus dimensions at some pre-decisional stage, referred to as "information-integration" categories, recruits posterior portions of striatum (i.e., the body and tail of caudate nucleus) that directly associate stimulus and response (e.g., Ashby et al., 1998; Ashby and Waldron, 1999; Ashby and Maddox, 2005). Because information-integration category input structures are designed so that no single dimension can independently signal the correct category membership, conscious effort to verbalize or explicit attempts to reason about the categorization decision are unhelpful, or even detrimental, to category learning (Ashby and Gott, 1988). Therefore, acquisition of information-integration categories becomes proceduralized instead of becoming reliant on working memory systems for explicit hypothesis-testing and allocation of executive attention to certain dimensions. This occurs via the posterior striatum such that direct associations between stimulus and response actions, implicitly acquired over the course of learning, are represented (Ashby et al., 1998; Yamamoto et al., 2013).

Both behavioral and neuroimaging findings have demonstrated that learning of information-integration categories recruits the direct stimulus-response association system associated with the posterior striatum to a greater extent than the explicit hypothesis-testing systems mediated by anterior striatum and the prefrontal cortex. In a behavioral study, Ashby et al. (2003) have found that switching stimulus-response key mappings in the course of training affected information-integration category learning, whereas explicit hypothesis-dependent category learning was unaffected. Similarly, compared to learning through variable response-category training (e.g., respond "yes" or "no" to "Is this A?" or "Is this B?"), consistent response mapping to stimulus category training (e.g., respond "A" or "B" to "Is this A or B?") was more advantageous for information-integration category learning (Maddox et al., 2004). In addition, manipulations known to recruit explicit attention/working memory systems, such as variations in the amount of information or the temporal delay in the feedback, hamper learning of information-integration categories (e.g., Maddox et al., 2003, 2008). Functional neuroimaging studies have also found that information-integration visual category learning induces activation in the posterior striatum as well as in lateral occipital and inferior temporal areas to a greater extent than explicit-verbal category learning (Seger and Cincotta, 2005). More specifically, Nomura et al. (2007) have observed learning-related activity in the body of the caudate nucleus for learning visual

Box 1 | Feedback-based "Reward-Prediction Error" Learning.



Feedback-based reward prediction error learning is driven by the outcome of feedback (e.g., reward) relative to the response. The reward prediction error (RPE) signal is generated based on the discrepancy between the actual feedback outcome that a learner receives and learner's expected feedback outcome. Learning proceeds as the discrepancy between the actual and expected outcome (i.e., RPE) decreases. Over the course of learning, one continues to learn correct responses that lead to rewarding outcomes in a given context, and the connection strengths among the input (bottom layer), perceptual (middle layer), and category response (top layer) layers changes according to the RPE magnitude in order to achieve rewarding outcomes in subsequent trials.

(A) Traditional explicit feedback-based tasks generate a RPE signal directed towards a specific perceptual domain related to a given explicit task. Learners' goal in these tasks is directly linked to correct categorization of a given sound stimulus. Learners are aware that outcome is directly related to categorization of an auditory signal and the RPE signal modulates representations of the task-relevant auditory perceptual domain. This type of learning can direct learners' attention to auditory stimuli and engage in explicit attempts to discover specific acoustic features defining category membership.

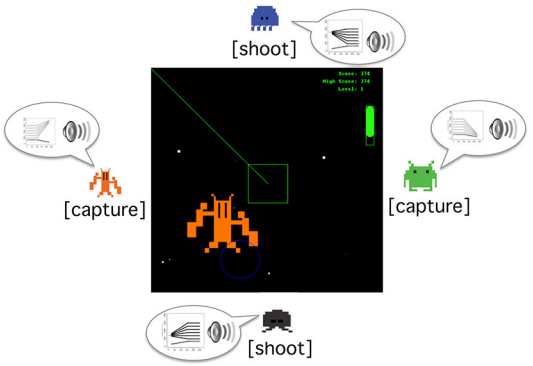
(B) Incidental training such as the videogame task may generate an RPE signal that propagates to multiple perceptual domains that support task success. In this type of tasks, learners have goals that are not directed to sound categorization, but to other features in the task (e.g., correct game actions on visual alien characters) that incidentally promote sound category learning. Outcome is linked to success in the game and learners are not aware of the relationship between outcome and sound categorization. Therefore, the RPE signal generated during learning may modulate auditory representations indirectly.

information-integration categories. These studies provide direct evidence that learning of visual categories requiring integration of multiple dimensions is mediated by a qualitatively different system than learning declarative, explicit knowledge that directs attention toward specific stimulus features. This may further suggest that optimal learning of procedural knowledge about categories may be achieved by learning of direct stimulus-response associations via recruitment of the posterior portion of the striatum.

Learning visual information-integration categories has close resemblance to the acquisition of speech sound categories (Chandrasekaran et al., 2014) due to the highly multidimensional nature of speech categories. This suggests that training paradigms that model aspects of the natural environment, and which do not involve explicit speech sound categorization judgments and that discourage active attempts to reason about the category mappings, may be more effective than directed speech categorization training. Evidence supporting this point of view comes from several studies that have examined incidental auditory and speech category learning in the context of a videogame training paradigm (Wade and Holt, 2005; Leech et al., 2009; Lim and Holt, 2011; Liu and Holt, 2011) (**Box 2**). Unlike explicit feedback-based categorization tasks, the videogame task incorporates a number of characteristics that mimic, and perhaps amplify, relationships among advantageous cues available in natural learning environments. Participants encounter rich correlations of multimodal cues (i.e., consistent auditory-category to visual-object pairing) while navigating a virtual space-themed gaming environment. The game encourages functional use of sound categories because the categories signal which alien creature is approaching and thereby reveal the appropriate action to take. Feedback arrives in the form of success or failure in executing these actions (capturing or shooting the aliens), rather than explicit feedback about the correctness of an overt categorization response. Even without overt categorization of sounds or directed attention to the sounds, listeners exhibit robust learning of multidimensional, artificial nonspeech sound categories (Wade and Holt, 2005). Furthermore, the videogame training with these nonspeech sounds induces learning-related neural changes that mimic those observed in speech categories learning (Leech et al., 2009; Liu and Holt, 2011). This method of auditory categorization training is also effective for non-native speech category learning. Just 2.5 h of game training with non-native speech sounds evokes non-native speech category learning comparable to traditional laboratory training involving overt categorization and explicit feedback across 2–4 weeks (Lim and Holt, 2011). These findings suggest that aspects of the videogame task may effectively engage learning mechanisms useful for acquiring sound categories.

A significant element of this training may be participants' motivation to successfully navigate the videogame and execute capturing and shooting actions. Since these actions are not

Box 2 | Videogame Training Paradigm (Wade and Holt, 2005).



In this game environment, each alien has a distinct color and shape, and it appears from a designated quadrant of the screen (as shown by positioning of the aliens). Correct game actions (e.g., shooting and capturing actions) are associated with an alien identity. In addition, each alien is associated with a particular category of sounds. When an alien appears on the game screen, an exemplar from its associated sound category is presented. Thus, the game training presents a complex and naturalistic learning environment, in which participants experience a rich correlation of multimodal (consistent pairing of auditory, visual, and motor) cues.

Participants' goal in the game task is to nativate and accurately aim at an alien in order to take an appropriate game action associated with the alien creature. Therefore, this training does not require directed attention to sounds or involve explicit categorization of sounds. Feedback is presented based on the success or failure of achieving goals in the game, not on the correctness of sound categorization.

The use of sound category information is gradually encouraged over the course of game play. The game becomes increasingly challenging as aliens appear further out from the center of the screen such that participants hear sounds before seeing an alien and the time window for action within each trial becomes shorter. Therefore, at high levels of the game, participants must rely on sound categorization to make correct game actions. The game encourages functional use of sound categories to accomplish the goals in the game.

directed at sound categorization *per se*, the videogame training paradigm may elicit internally-generated reward prediction error feedback signals from the basal ganglia that indirectly induce changes in sound category representations that correlate to the success in the task (**Box 1B**). Processing task-relevant rewards incidentally in relation to sound categories may inhibit explicit attention to sounds, which can actually discourage perceptual learning (Tsushima et al., 2008; Gutnisky et al., 2009). Moreover, the increased engagement imposed by the game task requires faster execution of navigation and action responses. This task demand may distract individuals from making explicit hypotheses about specific acoustic features related to category mapping and, in turn, motivate learning automatic responses. Therefore, the Wade and Holt (2005) videogame may provide a training environment better-suited to recruiting the posterior striatal system that has been implicated in the learning of information-integration categories, as compared to directed categorization tasks. Supporting this possibility, we have found sound category learning within the videogame paradigm engages the posterior striatum (i.e., the caudate body) (Lim et al., 2013), which may contribute to learning-related perceptual plasticity (see Tricomi et al., 2006, discussion). This may explain the relative effectiveness of non-native speech category learning observed in the videogame (Lim and Holt, 2011), as compared to directed speech categorization training. These findings suggest that the basal ganglia play a role in learning within the Wade and Holt videogame task, and that its recruitment might be significant in supporting changes in cortical representations of the to-be-learned sound categories.

Another recent speech category learning study has emphasized the crucial role of reward-driven striatal-learning systems in non-native speech category learning. This study directly applied findings from the visual category learning literature (see Ashby and Maddox, 2005, for a review), which supports the existence of differential striatal learning systems recruited via principled manipulations to task structure and stimulus input distributions. By manipulating the schedule and content of trial-by-trial feedback, Chandrasekaran et al. (2014) have found that the extent of non-native speech category learning is greater in training tasks that tap into striatum-dependent procedural learning as compared to explicit hypothesis-testing learning. More specifically, compared to delayed feedback, immediate feedback occurring within 500 ms after a response can induce learning. This is hypothesized to occur because the 500-ms window aligns with the timecourse of influence of dopamine signals from feedback. Within this window, a brief dopamine signal can effectively influence cortico-striatal synapses for processing a stimulus and response while they remain active, which may enable learning of direct stimulus-response associations (see Ashby et al., 2007, for a review). Likewise, minimal information in the feedback (e.g., correct vs. incorrect) without information about the correct category mapping may minimize the chance of recruitment of the explicit hypothesis-testing process, and lead to greater engagement of the striatum-dependent procedural learning. Like the Wade and Holt (2005) videogame, this study also demonstrates that the nature of the task (in Chandrasekaran et al., 2014 the timing of feedback presentation) may modulate the recruitment of striatum-mediated learning, which can subsequently affect the outcome of non-native speech category learning.

Similarly, another line of research has demonstrated the effectiveness of implicit over explicit training procedures for perceptual learning. In studies of visual perceptual learning, some investigations have emphasized the role of diffuse reinforcement signals (specifically, dopaminergic reinforcement signals) in inducing perceptual plasticity and learning regardless of the direct relevance to the perceptual stimuli used in the task (Seitz and Watanabe, 2003, 2005, 2009; Seitz et al., 2009). Directly applying this paradigm, Vlahou et al. (2012) has shown that implicit, reward-contingent exposure of to-be-learned non-native speech stimuli seems to be more advantageous than explicit feedback-based exposure. Although this line of work has not implicated the striatum in learning, it has demonstrated the

advantage of reward signals and of implicit vs. explicit training tasks for learning speech.

Overall, these results suggest that understanding the task demands and stimulus characteristics that effectively recruit the basal ganglia learning system can reveal approaches to promoting adult speech category learning. Regardless of whether the training paradigm involves overt, experimenter-provided feedback as in directed categorization tasks or indirect feedback as in the videogame task, the basal ganglia play a role in promoting learning based on outcome feedback. Significantly, however, differences in task characteristics may have important consequences for the manner by which learning is achieved (**Box 1**) inasmuch as they engage distinct basal ganglia-thalamo-cortical loops. Overt, category learning tasks that provide feedback about the accuracy of a speech category judgment may promote learning by directing explicit attention to sounds to discover critical stimulus characteristics relevant to category membership (Logan et al., 1991; Francis and Nusbaum, 2002; Heald and Nusbaum, 2014). Learning of explicit goal-directed actions based on feedback appears to be mediated by the anterior portion of the dorsal striatum, which interacts with executive and attention/working memory systems.

On the contrary, training tasks that recruit the posterior striatum may be advantageous for promoting optimal non-native speech category learning, because they may bypass an explicit hypothesis-testing system involving the anterior striatum, and instead promote a form of procedural learning that is more suited for learning categories with an information-integration structure, including speech categories (Chandrasekaran et al., 2014). One possible advantage of posterior striatum recruitment in category learning is that it can interact with sensory cortex to a greater extent than the anterior striatum, for which interaction with sensory cortex is mediated through the frontal cortex. Learning of implicit stimulus-action relationships appears to involve striatal regions in the posterior striatum, which are known to develop automatic responses based on consistent reward experiences (Seger and Cincotta, 2005; Cincotta and Seger, 2007; Kim and Hikosaka, 2013; Yamamoto et al., 2013), thereby prohibiting the use of non-optimal strategies for categorization. Therefore, the Wade and Holt (2005) videogame task may indirectly promote learning of sound category features even as listeners' attention is directed away from the sounds and toward other task goals, such as making correct game actions to respond to the visual aliens. The task demands of the primary task (navigating the videogame, for example) may be time and resource demanding enough to discourage active attempts to reason about category-diagnostic dimensions. Or, learners might be truly unaware that the outcomes of their actions are linked to the learning of category-relevant features. Future investigations are needed to clarify the role of the posterior striatum in category learning, specifically regarding the mechanisms by which category learning is actually achieved and the nature of learned categories represented in the posterior striatum.

BASAL GANGLIA INTERACTIONS WITH SENSORY CORTEX

Previous neuroimaging studies involving auditory category learning have shown that category learning can change cortical processing for the learned sounds. In particular, the observed effect of feedback valence on the activation of the auditory regions in the superior temporal gyrus (Tricomi et al., 2006) may suggest that processing of feedback information via the basal ganglia can induce changes in the sensory cortical regions for learned phonetic representations. For example, incidental learning of nonspeech sound categories within the Wade and Holt (2005) videogame recruits posterior superior temporal sulcus (pSTS) regions associated with speech processing in response to the newly-acquired nonspeech categories (Leech et al., 2009). This change may be occurring at an early processing stage, as the same category learning can elicit changes in the evoked response potential within 100-ms after the onset of the learned sounds (Liu and Holt, 2011). Furthermore, explicit feedback-based training of sound categories has been shown to promote activity changes in the auditory cortical regions, such that they respond in a categorical fashion (e.g., Callan et al., 2003; Golestani and Zatorre, 2004; Dehaene-Lambertz et al., 2005; Desai et al., 2008; Liebenthal et al., 2010; Lee et al., 2012; Ley et al., 2012). The observed learning-related changes of sensory cortical processing suggests that the sensory cortex is affected by "teaching signals" elicited from training (e.g., reward-based learning signals based on feedback). The basal ganglia may support such interaction with the sensory regions.

As noted earlier, the basal ganglia are known to have multiple anatomical cortico-striatal loops that innervate widespread areas of the cerebral cortex, including motor, cognitive and perceptual regions (see Alexander et al., 1986, for a review). These loops are organized in a topographical manner such that information in each loop projects to specific regions in the striatum and in the thalamus. This information is subsequently fed back to distinct cortical regions (Parent and Hazrati, 1995) via "closed loops," which send reciprocal projections to the originating cortical regions (Alexander et al., 1986) and "open loops," which ultimately terminate at different cortical regions (Joel and Weiner, 1994). These anatomical loops serve distinct functions, the nature of which depends on the pattern of cortical projections. Among these multiple cortico-striatal loops, the visual loop from inferior temporal regions of cerebral cortex has been commonly implicated in perceptual category learning (see Seger, 2013, for a review; **Figure 2**). Although auditory regions in the superior temporal region form cortico-striatal projections similar to the visual loop, the auditory loop has been relatively less studied. Therefore, we first focus on the findings from the visual cortico-striatal loop, which would be relevant for understanding the role of the auditory cortico-striatal loop inasmuch as they reveal how posterior sites of basal ganglia may influence sensory cortical processing.

The presence of the visual cortico-striatal loop indicates that the striatum is able to interact with cortical regions responsible for sensory processing. Animal neurophysiology studies have demonstrated that the body and tail of the caudate nucleus contain neurons that respond to visual input. Studies examining the function of this visual loop have shown that animals with specific lesions in the tail of the caudate are impaired in visual discrimination learning (Packard et al., 1989; Packard and McGaugh, 1992). Another study has shown that among all connections from the visual cortex, only connections between the inferior temporal

cortex and the striatum are necessary and sufficient to achieve visual discrimination learning (Gaffan and Eacott, 1995).

Human neuropsychological and neuroimaging studies have provided converging evidence to support the role of the striatum in visual category learning. Studies have shown that Parkinson's and Huntington's disease patients are impaired in learning visual categories that require information integration (Filoteo et al., 2001; Ashby and Maddox, 2005). Human fMRI studies have demonstrated recruitment of the body and tail of caudate nucleus during visual categorization (Cincotta and Seger, 2007; Nomura et al., 2007). These converging findings from both animal and human research demonstrate the role of the striatum (specifically, the body and tail of the caudate nucleus) in category learning within the domain of visual perception. Based on the fact that reward-related learning within the striatum can modulate synaptic efficacy across relevant cortico-striatal loops (Houk and Wise, 1995), the striatum might play a significant role in inducing learning-related representational changes in visual cortex.

It is of note that striatal-mediated visual category learning research has mostly focused on "open loop" projections of cortico-striatal pathways. Research typically has assumed that perceptual representations are computed and selected by the visual cortex whereas the striatum is responsible for selecting an appropriate category decision, which is then transmitted to motor cortex to execute a response (Ashby et al., 1998; Ashby and Waldron, 1999; Ashby and Spiering, 2004). In other words, most research has been directed at how basal ganglia-dependent circuits acquire information that can be used to guide "action selection" in response to a visual stimulus (see Seger, 2008, for a review). Therefore, these studies have often been concerned with interactions among different cortico-striatal loops: projections from the sensory regions (i.e., high-level visual regions) to the striatum, and projections from the striatum to frontal or motor cortical regions (Lopez-Paniagua and Seger, 2011). In contrast, relatively less attention has been directed to the role of the "closed" striatal projection back to visual cortex (or sensory cortex, in general). An animal viral tracing study has shown that the basal ganglia system indeed projects back to the inferior temporal cortex (Middleton and Strick, 1996), the high-level visual cortical region that plays a critical role in visual recognition and discrimination (Mishkin, 1982; Ungerleider and Mishkin, 1982) and visuomotor associations (Mishkin et al., 1984). In humans, damage to the visual loop striatal circuitry has been associated with deficits in face perception (Jacobs et al., 1995). This evidence indicates that the striatum has the capacity to influence sensory processing within visual cortex.

The striatum may affect visual processing through dopamine-dependent synaptic plasticity within the basal ganglia (Kerr and Wickens, 2001; Centonze et al., 2003; Calabresi et al., 2007). A neurocomputational model proposed by Silkis (2007, 2008) shows that reorganization of the synaptic network via dopamine can differentially modulate the efficiency of strong and weak cortico-striatal inputs in a manner analogous to the basal ganglia's role in action selection. When strong visual cortico-striatal input occurs simultaneously with dopamine release, the basal ganglia circuit can be reorganized to ultimately disinhibit the visual cortical neurons that were strongly activated, and conversely inhibit neurons that were weakly activated. Therefore, if either top-down or bottom-up visual attention can evoke dopamine release (Kähkönen et al., 2001), the cortico-basal ganglia network may be reorganized to affect processing that occurs within visual regions. Through this type of mechanism, feedback-based dopaminergic reinforcement signals from the training experience could affect sensory processing regions via the basal ganglia. In support of this argument, dopamine release associated with the receipt of reward can affect early sensory/perceptual processing. Incidental delivery of reward during passive viewing of visual stimuli has been shown to induce changes in low-level visual discrimination. Perceptual sensitivity is selectively increased to process features of a stimulus that were simultaneously presented with reward, whereas there was no change in sensitivity to process unrewarded stimuli features (Seitz and Watanabe, 2003, 2009; Seitz et al., 2009).

Another possible mechanism by which the striatum could interact with sensory cortex is via the prefrontal cortex. As noted in section Overview of the Basal Ganglia and Reinforcement Learning, the basal ganglia effectively learn stimulus-action-outcome associations leading to rewards via dopamine release. This reward-related stimulus-action representation may reside in frontal higher-order cognitive or motor regions. Across various learning studies, the prefrontal cortex is known to represent "goal-directed" actions in response to a given stimulus (Petrides, 1985; Wallis et al., 2001; Muhammad et al., 2006). It has been proposed that this learning in the prefrontal cortex is achieved through recurrent interaction with the basal ganglia; reward-driven stimulus-response associations rapidly acquired by the basal ganglia are projected to the prefrontal cortex through a cortico-striatal loop, while the prefrontal cortex slowly integrates and binds multiple information sources to build higher-order representations (i.e., the process of generalization) (Pasupathy and Miller, 2005; Miller and Buschman, 2008). Therefore, in the context of category learning, the basal ganglia may induce a "goal-directed" representation of appropriate category response toward a given stimulus in the prefrontal cortex (Kim and Shadlen, 1999; Freedman et al., 2001; McNamee et al., 2013), which in turn may exert top-down attentional modulation on sensory regions to selectively respond to learning-relevant sensory information (Duncan et al., 1997; Desimone, 1998). It remains unclear whether the frontal cortex exerts a direct influence on the sensory regions or whether top-down attention modulates plasticity of the cortico-basal ganglia-thalamic circuit via dopamine release (see Miller et al., 2011, discussion; Skinner and Yingling, 1976; Silkis, 2007). Either possibility invites consideration of the role of the basal ganglia in indirectly or directly modulating attention (van Schouwenburg et al., 2010), which can ultimately tune sensory cortex to form robust category representations (Fuster et al., 1985; Beck and Kastner, 2009) and to exhibit experience- and learning-dependent neural response selectivity to category-relevant over category-irrelevant sensory features (e.g., Sigala and Logothetis, 2002; Op de Beeck et al., 2006; Folstein et al., 2013; van der Linden et al., 2014).

These loops provide a means by which the striatum can interact with sensory cortical regions and may indicate a role for the basal ganglia in auditory/speech category learning. Compared to the role of visual cortico-striatal loop, relatively

less is known about auditory cortico-striatal loop that links auditory cortical regions and the basal ganglia. Nevertheless, animal neurophysiological research has shown a direct link between the striatum and auditory cortex, which strongly implies the presence of an auditory cortico-striatal loop. Within the body of the caudate, auditory cortex projections converge onto a region that is distinct from the striatal site receiving cortical projections from visual processing regions (Arnauld et al., 1996). The sector of the striatum that receives auditory cortical projections projects back to the auditory cortex via the output structures of the basal ganglia (Parent et al., 1981; Moriizumi et al., 1988; Moriizumi and Hattori, 1992; see Parent and Hazrati, 1995, for a review). Non-human primate neurophysiology studies also have demonstrated that different auditory cortex regions (i.e., primary, secondary) form connections with different sectors of the striatum (Van Hoesen et al., 1981; Yeterian and Pandya, 1998). Importantly, a recent study has demonstrated in rats that auditory cortico-striatal projections influence behavioral performance during a reward-based frequency discrimination task (Znamenskiy and Zador, 2013).

There is also emerging evidence from human neuroimaging revealing the role of the auditory cortico-striatal loop. Geiser et al. (2012) have shown that recruitment of a cortico-striatal system facilitates auditory perceptual processing in auditory temporal cortex. Directly relevant in the context of learning speech categories, Tricomi et al. (2006) observed that observed recruitment of the striatum among native Japanese adults learning of English /r/ and /l/ categories via an overt categorization task with feedback. This study demonstrated a possible interaction between striatum system and the auditory cortex, such that differential activity was observed in the caudate nucleus as well as in the left superior temporal gyrus, a cortical region known to be associated with non-native phonetic learning (Callan et al., 2003; Golestani and Zatorre, 2004), across correct vs. incorrect trials. Although it is still unclear whether the recruitment of the striatum in the overt categorization task involves the top-down influence from the higher-order cortical regions (e.g., frontal cortex) or a direct influence from the striatum to auditory regions, this evidence may indicate that the striatum, recruited by feedback-based training tasks, interacts with cortical regions processing speech. This striatal innervation in learning may effectively induce learning-related plasticity, which may ultimately influence cortical representations of the newly learned non-native speech categories.

In addition to the striatal interaction with the auditory processing regions via the "closed" auditory loop, the "open loop" pathway of the basal ganglia to frontal and motor regions may contribute to speech category learning by facilitating sensory and motor interactions. Previous neuroimaging studies investigating speech perception have demonstrated interactions between the speech perception and production (i.e., sensory and motor interactions). For example, listening to speech sounds activates both auditory regions (i.e., superior temporal cortex) and motor regions involved in speech production (e.g., Wilson et al., 2004; Wilson and Iacoboni, 2006). Perception of distinct speech categories is reflected in neural activity patterns in the frontal and motor regions including Broca's area and pre-supplmentary motor area (pre-SMA), known to participate in speech motor planning and articulatory processing (Lee et al., 2012). Moreover, learning non-native speech categories has also been shown to engage similar regions in the frontal and motor areas (Callan et al., 2003; Golestani and Zatorre, 2004), which interact with the basal ganglia via cortico-striatal loops (Alexander et al., 1986; Middleton and Strick, 2000; Clower et al., 2005). Although the nature of the speech perception and production link (see Lotto et al., 2009, for a review) and its role in speech category acquisition are yet to be discovered, the basal ganglia's closed and open loop projections have the potential to facilitate learning of speech categories via interactions between perception- and action-related representations of speech categories.

CATEGORY GENERALIZATION THROUGH CONVERGENCE OF THE BASAL GANGLIA

Previous studies investigating basal ganglia-mediated category learning have emphasized the learning of representations at the level of category decision-making to trained exemplars (e.g., Ashby et al., 1998). Therefore, it remains uncertain whether the basal ganglia contribute to forming perceptual category representations that are generalizable across variable instances of a class (Palmeri and Gauthier, 2004). This is an important issue for speech category learning, as generalization of learning to new exemplars is a hallmark of categorization. Although there might be multiple factors that can contribute to generalization (e.g., attentional modulation), the basal ganglia may play a crucial role.

Cortical information funnels through the basal ganglia via multiple cortico-striatal loops. Massive projections from widespread cortical areas are reduced as they reach the striatum and globus pallidus. The number of neurons from cortex to the striatum is reduced on the order of 10 (Zheng and Wilson, 2002), which is further reduced at the globus pallidus on the order of 10^2–10^3 (Percheron et al., 1994), thereby creating a highly convergent "funneling" of information within the basal ganglia (Flaherty and Graybiel, 1994). With this convergence of cortical input to the basal ganglia approximately at a ratio of 10,000:1 (Wilson, 1995), compressed cortical information is fed back to the cortical regions that send projections to the striatum via basal ganglia output.

The exact degree and the pattern of this convergence have been under debate. Initially, the cortex was thought to innervate the striatum in a topographical fashion such that a group of spatially adjacent cortical input would project to a localized region within the striatum (Webster, 1961), thus removing redundancy of the input. However, the later findings have shown that the striatum is innervated by distributed, yet inhomogeneous, cortical input (Selemon and Goldman-Rakic, 1985; Malachi and Graybiel, 1986), whereby the striatum acts as a "pattern detector" across cortical input (Zheng and Wilson, 2002; Bar-Gad et al., 2003). In other words, a specific pattern of cortical input even originating from spatially sparse cortical regions may be required to activate corresponding striatal neurons. In this way, the striatum may represent functional organization, rather than the spatial topography of the cortex (e.g., Flaherty and Graybiel, 1993, 1994). Although such a pattern of innervation can raise questions about the extent of convergence, the compression of cortical information within

the striatum is inevitable. With the reduced number of striatal neurons, the striatum cannot represent all possible patterns of cortical input (Zheng and Wilson, 2002). This constraint allows the basal ganglia to reduce or compress cortical information, which is eventually fed back to the cortex.

This converging characteristic of the basal ganglia might be quite suitable for generalization by preserving learning-relevant information and diminishing stimulus-specific information. The computational model by Bar-Gad et al. (2003) illustrates this dimension reduction mechanism of the basal ganglia; as information is reduced, reward-related information is retained and enhanced whereas non-rewarded information is inhibited or unencoded. This computational scheme could be useful for forming category representations capable of producing generalization across variable instances by strengthening category-relevant over -irrelevant information within sensory cortex, via recurrent projections with the basal ganglia.

The basal ganglia's potential role in information reduction could provide a useful and important neural mechanism for the facilitation of perceptual category learning. Across visual and auditory domains, perceptual category learning studies have emphasized the importance of stimulus variability in acquiring robust and "generalizable" category formation. Posner and Keele (1968) have observed that training with high-variability stimuli during visual pattern classification task is more advantageous than training with low-variability stimuli, as assessed by the ability to generalize learning to accurately classify novel visual patterns. Similarly in the domain of speech category learning, studies have emphasized the benefits of high-variability in training stimuli (with speech from multiple talkers, and speech contexts, e.g., Logan et al., 1991; Lively et al., 1993, 1994) as training with low-variability fails to generalize listeners' learning to novel sounds. There is a perceptual cost associated with learning categories from multi-speaker stimuli as it can lead to increased response times and reduced overall categorization accuracy (Mullennix et al., 1989). Nevertheless, training with low-variability (e.g., single-speaker's speech) stimuli may lead to non-optimal category learning dependent on information diagnostic to that speaker's speech, while training with multi-speaker stimuli can highlight category-relevant acoustic cues. Because highly variable stimulus input can create enough variance in category-irrelevant dimensions, learners may selectively encode less-variable, but category-relevant dimensions to form representations that effectively capture the information most diagnostic of category membership (Lively et al., 1993; see Pisoni, 1992), which can be applied upon encountering novel instances. The mechanism of high-variability training promoting perceptual category learning has a close resemblance to the basal ganglia's potential role in input dimension-reduction.

The dimension reduction characteristic of the basal ganglia may serve a beneficial role in natural speech category learning. A main challenge of speech perception/categorization is parsing highly variable acoustic signals as linguistically-relevant units (see Holt and Lotto, 2010, for a review). As mentioned above, speech is inherently multidimensional such that many acoustic cues can be used to determine category membership. However, it is important to note that although multiple cues covary with speech category identity, not all acoustic cues are equally weighted for perception; listeners rely on certain acoustic dimensions more heavily than others for categorization (Francis et al., 2000; Idemaru et al., 2012). Based on the distributional characteristics of speech categories in a given language, listeners learn to rely more on acoustic dimensions that are most diagnostic of category membership. Of course, there might be an accumulation of experience with statistical regularity of the speech category input (i.e., similarity across exemplars within a category; see computational models by McMurray et al., 2009; Toscano and McMurray, 2010). Nevertheless, there appears to be a prioritizing of category-relevant dimensions in speech perception. The mechanism of information reduction via cortico-striatal convergence may serve a supportive role for facilitating extraction of critical and behaviorally significant information relevant for categorization. This mechanism may give rise to robust perceptual representations.

GENERAL CONCERNS AND FUTURE DIRECTIONS

LEARNING-RELATED REPRESENTATIONS

It is of note that there exist discrepancies among independent lines of research in perceptual category learning and basal ganglia-mediated category learning research. General perceptual category and object learning studies have been concerned largely with observations of learning-related neural changes in the sensory cortices as an outcome of learning. Perception (and sensory cortex) is tuned to exhibit a selective improvement in processing category-relevant over -irrelevant dimensions (Goldstone, 1994; Gureckis and Goldstone, 2008). In contrast, basal ganglia-mediated category learning research has mostly been concerned with issues regarding how perceptual categories are acquired, with the presumption that learning-related representational change occurs at the level of action selection and decision making about a given category instance (i.e., associations between a stimulus and a correct categorization response), leaving sensory representations relatively unaffected (e.g., Ashby et al., 1998; Ashby and Waldron, 1999; Ashby and Spiering, 2004). Because of this orientation, previous studies have indicated the basal ganglia in category learning regardless of the presence of category structure. These studies have not differentiated or directly compared the process of learning structured categories that require integration of multiple dimensions vs. arbitrary/unstructured category exemplars randomly distributed without any specific category boundaries (Seger and Cincotta, 2005; Cincotta and Seger, 2007; Seger et al., 2010; Lopez-Paniagua and Seger, 2011; Crossley et al., 2012), although different category input distributions can have a notable impact on sensory processing and learning (Wade and Holt, 2005; Holt and Lotto, 2006; Lim et al., 2013).

A similar tension exists in interpreting results of perceptual category learning studies. Some studies have demonstrated neural changes in sensory regions after learning (e.g., Sigala and Logothetis, 2002; Guenther et al., 2004; Desai et al., 2008; Ley et al., 2012; van der Linden et al., 2014), even when listeners are passively exposed to learned category instances after training (Leech et al., 2009; Liu and Holt, 2011). On the contrary, instead of sensory regions, other studies have suggested that learned categories and objects are represented in the higher-order cortical

areas like frontal regions (e.g., Freedman et al., 2001, 2003; Jiang et al., 2007). This view is in line with basal ganglia-mediated category learning research that posits that the learning-related representational change occurs only at the level of action selection and decision-making. As such, the target of category-learning representational change is as yet unknown. However, it is important to acknowledge that that learning-related plasticity arising either in sensory cortical processing or other decision-related cortical regions may depend critically on how perceptual categories are defined (Folstein et al., 2012) and the tasks by which they are learned.

Future research will be needed to resolve whether category learning is better conceived of as change in decision mapping vs. sensory perception and to determine whether both types of representational change may be simultaneously developed over the course of learning via multiple cortico-striatal loops. This possibility would lead to learned stimulus-response associations to strengthen the behavioral significance of perceptual representations, which perhaps could induce changes in the sensory-level processing to selectively enhance perception of category-diagnostic features.

NATURALISTIC LEARNING ENVIRONMENTS FOR SPEECH

Although the basal ganglia have been implicated in visual category learning, their role has been rarely considered in understanding speech category learning. The discussion above highlights some reasons to believe that characteristics of basal ganglia function may support second-language speech category learning under the right task demands. An open question is whether this system might support first-language speech category learning. Infants fairly rapidly attune to the distributional regularities of native language speech categories without explicit instruction (e.g., Aslin et al., 1998; Maye et al., 2002). A common notion is thus that infants acquire native speech categories without feedback, perhaps through mechanisms related to statistical learning (see Kuhl, 2004, for a review). Since infants exhibit statistical learning in passive listening laboratory tasks (e.g., Saffran et al., 1996, 1999; Aslin et al., 1998; Maye et al., 2002), other learning mechanisms have not been widely considered.

However, an important concern is whether the learning systems engaged by passive laboratory tasks would scale up to accommodate the complexity of natural language learning environments. In a natural listening environment, listeners experience highly acoustically-variable phonemic sounds in fluent and continuous speech rather than as isolated instances. This adds the additional challenge of learning the perceptual mapping of sound to functionally equivalent language-specific units (such as phonemes, or words) while simultaneously parsing continuous speech input. In addition, speech exposure often occurs within complex visual scenes for which there are multiple potential referents, creating additional learning challenges (Medina et al., 2011). This complexity introduces an explosion of potentially-relevant statistical regularities, leading some to suggest that passive computation of statistics in the speech input alone cannot induce early speech learning within complex natural speech settings (Kuhl, 2007). Evidence suggests that statistical learning within natural language environments may be supported by modulation from attentional and motivational factors (Kuhl, 2003; Kuhl et al., 2003; Toro et al., 2005), contingent extrinsic reinforcers like social cues (Goldstein et al., 2003; Gros-Louis et al., 2006), and the presence of correlated multimodal (e.g., visual) inputs (Hollich et al., 2005; Teinonen et al., 2008; Yeung and Werker, 2009; Thiessen, 2010). Similar to the learning process engaged by the videogame training, the indirect influence of such signals on early speech processing may indicate a potential role for recruitment of the basal ganglia learning system that incidentally facilitates acquisition of native speech categories. Investigating this further in future research will help to refine models of first-language speech category acquisition.

A different line of research has suggested that implicit, task-irrelevant perceptual features of rewarded stimuli can be learned with passive exposure via a diffuse dopamine signal (Seitz and Watanabe, 2003, 2005; Seitz et al., 2010). Although this line of research has not implicated the specific role of the striatum, Vlahou et al. (2012) demonstrates the importance of reward-related learning signals on perceptual plasticity (Seitz et al., 2009) useful for non-native speech category learning. However, it is of note that the task-irrelevant training paradigm does not have any component to signal information about the functional distinctiveness across different categories or to induce reward or dopamine signals throughout learning, except for the external rewards that are implicitly paired with the stimuli by the experimenter. This task-irrelevant perceptual learning may lead to perceptual attunement to very specific stimulus information that coincides with external reward delivery. Due to such specificity, non-native speech learning in this task seems to be limited to familiar training speech sounds that have been paired with external rewards and does not generalize to novel sound stimuli (Vlahou et al., 2012). Although the thresholds of non-native speech sound discriminability change as a result of this training, it is not yet known whether task-irrelevant perceptual learning can lead to perceptual *category* learning and generalization. Nonetheless, although research on task-irrelevant perceptual learning does not yet converge with the learning challenges of non-native speech category learning, it does provide insight in the learning systems that may be engaged to modify sound perception. It may be fruitful to try to bridge this gap in future research.

The Wade and Holt (2005) videogame training paradigm described above also falls short in modeling the naturalistic learning environment for learning speech categories. However, it does provide a means of manipulating signals influential in first language speech category acquisition such as motivational factors, contingent reinforcement, and multimodal correlations. It also presents the possibility of scaling up the learning challenges. In recent research Lim et al. (under review) have found that adults can discover non-native speech and also nonspeech sound categories from continuous, fluent sound input in the context of the Wade and Holt (2005) videogame. This learning generalized to novel exemplars, indicative of robust category learning. Given that research implicates the basal ganglia in learning within this task (Lim et al., 2013), there is the opportunity for future research to compare and contrast basal ganglia-mediated learning with that arising from passive learning.

CONCLUSION

The basal ganglia are a very complex and intricate neural structure, consisting of multiple sub-structures that interact with most cortical areas through diverse connections. The structure has been highly implicated in motor functions. However, general learning studies outside of the speech/auditory domain have revealed its contribution to cognitive functions, particularly in learning from external feedback to form goal-directed and procedural behaviors as well as learning visual categories.

In the domain of speech category learning and elsewhere, research commonly uses explicit feedback-based tasks to induce effective learning. Although this type of task engages the basal ganglia system during learning, and is known to be effective for acquisition of non-native speech categories (McCandliss et al., 2002; Tricomi et al., 2006), speech learning studies have put relatively less emphasis on the nature of the training experience influencing the learning process and outcome. Likewise, existing neurobiological and computational models of speech processing (e.g., the dual-stream neural account of Hickok and Poeppel, 2004; or the TRACE computational model of McClelland and Elman, 1986, but see Guenther, 1995) have focused on cortical networks and have not widely considered how subcortical structures like the basal ganglia participate in speech category acquisition or captured more than limited forms of learning. Although it has great relevance, current theories do not address the role of different training experiences on recruiting the basal ganglia and the corresponding effects on behavioral and neural changes for speech perception and learning. Therefore, a better understanding of learning-related functions of the basal ganglia system may be important in elucidating how effective speech category learning occurs. This may have rich benefits for optimizing training environments to promote perceptual plasticity in adulthood. Furthermore, understanding of the basal ganglia system may provide a broader understanding of language learning in general as it has been implicated in various aspects of language-related processing (Ullman et al., 1997; Doupe and Kuhl, 1999; Kotz et al., 2009).

The topics of speech perception and learning, and basal ganglia-mediated category learning, have been largely studied independently. Speech perception, once considered a "special" perceptual system, has only recently begun to be studied in a manner that fully incorporates general cognitive/perceptual learning research on the development of perceptual representations. On the other hand, studies of basal ganglia function with regard to category learning have emphasized understanding of the process of learning category-relevant decisions rather than learning-related changes in perceptual organization. However, these separate lines of research share commonalities. We have attempted to argue that there is great potential in bridging efforts to understand speech perception and learning with general cognitive neuroscience approaches and neurobiological models of learning.

ACKNOWLEDGMENTS

This work was supported by training grants to Sung-Joo Lim from the National Science Foundation (DGE0549352), the National Institute of General Medical Sciences (T32GM081760), and the National Institute on Drug Abuse (5T90DA022761-07), grants to Lori L. Holt from the National Institutes of Health (R01DC004674) and the National Science Foundation (22166-1-1121357), and grants to Julie A. Fiez from the National Institute of Health (R01HD060388) and the National Science Foundation (SBE-0839229).

REFERENCES

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally linking basal ganglia and cortex. *Annu. Rev. Neurosci.* 9, 357–381. doi: 10.1146/annurev.ne.09.030186.002041

Amalric, M., and Koob, G. F. (1993). Functionally selective neurochemical afferents and efferents of the mesocorticolimbic and nigrostriatal dopamine system. *Prog. Brain Res.* 99, 209–226. doi: 10.1016/S0079-6123(08)61348-5

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *J. Phon.* 32, 233–250. doi: 10.1016/S0095-4470(03)00036-6

Arnauld, E., Jeantet, Y., Arsaut, J., and Demotes-Mainard, J. (1996). Involvement of the caudal striatum in auditory processing: c-fos response to cortical application of picrotoxin and to auditory stimulation. *Mol. Brain Res.* 41, 27–35. doi: 10.1016/0169-328X(96)00063-0

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* 105, 442–481. doi: 10.1037/0033-295X.105.3.442

Ashby, F. G., and Ell, S. W. (2001). The neurobiology of human category learning. *Trends Cogn. Sci.* 5, 204–210. doi: 10.1016/S1364-6613(00)01624-7

Ashby, F. G., Ell, S. W., and Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Mem. Cognit.* 31, 1114–1125. doi: 10.3758/BF03196132

Ashby, F. G., Ennis, J. M., and Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychol. Rev.* 114, 632–656. doi: 10.1037/0033-295X.114.3.632

Ashby, F. G., and Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* 14, 33–53. doi: 10.1037/0278-7393.14.1.33

Ashby, F. G., and Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.* 56, 149–178. doi: 10.1146/annurev.psych.56.091103.070217

Ashby, F. G., and O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends Cogn. Sci.* 9, 83–89. doi: 10.1016/j.tics.2004.12.003

Ashby, F. G., and Spiering, B. J. (2004). The neurobiology of category learning. *Behav. Cogn. Neurosci. Rev.* 3, 101–113. doi: 10.1177/1534582304270782

Ashby, F. G., and Waldron, E. M. (1999). On the nature of implicit categorization. *Psychon. Bull. Rev.* 6, 363–378. doi: 10.3758/BF03210826

Aslin, R. N., Jusczyk, P. W., and Pisoni, D. B. (1998). "Speech and auditory processing during infancy: constraints on and precursors to language," in *Handbook of Child Psychology*, Vol. 2, eds W. Damon, K. Kuhn, and R. S. Siegler (New York, NY: John Wiley & Sons), 147–198.

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog. Neurobiol.* 71, 439–473. doi: 10.1016/j.pneurobio.2003.12.001

Beck, D. M., and Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Res.* 49, 1154–1165. doi: 10.1016/j.visres.2008.07.012

Beninger, R. J. (1983). The role of dopamine in locomotor activity and learning. *Brain Res.* 287, 173–196. doi: 10.1016/0165-0173(83)90038-3

Berns, G. S., McClure, S. M., Pagnoni, G., and Montague, P. R. (2001). Predictability modulates human brain response to reward. *J. Neurosci.* 21, 2793–2798.

Best, C. T. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed W. Strange (Timonium, MD: York Press), 171–204.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *J. Acoust. Soc. Am.* 101, 2299–2310.

Calabresi, P., Picconi, B., Tozzi, A., and Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci.* 30, 211–219. doi: 10.1016/j.tins.2007.03.001

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., and Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage* 19, 113–124. doi: 10.1016/S1053-8119(03)00020-X

Centonze, D., Grande, C., Saulle, E., Martin, A. B., Gubellini, P., Pavón, N., et al. (2003). Distinct roles of D1 and D5 dopamine receptors in motor activity and striatal synaptic plasticity. *J. Neurosci.* 23, 8506–8512.

Chandrasekaran, B., Yi, H.-G., and Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychon. Bull. Rev.* 21, 488–495. doi: 10.3758/s13423-013-0501-5

Cincotta, C. M., and Seger, C. A. (2007). Dissociation between striatal regions while learning to categorize via feedback and via observation. *J. Cogn. Neurosci.* 19, 249–265. doi: 10.1162/jocn.2007.19.2.249

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108, 804–809. doi: 10.1016/j.cognition.2008.04.004

Clower, D. M., Dum, R. P., and Strick, P. L. (2005). Basal ganglia and cerebellar inputs to "AIP." *Cereb. Cortex* 15, 913–920. doi: 10.1093/cercor/bhh190

Crossley, M. J., Madsen, N. R., and Ashby, F. G. (2012). Procedural learning of unstructured categories. *Psychon. Bull. Rev.* 19, 1202–1209. doi: 10.3758/s13423-012-0312-0

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711. doi: 10.1038/nn1560

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., and Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *Neuroimage* 24, 21–33. doi: 10.1016/j.neuroimage.2004.09.039

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., and Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* 84, 3072–3077.

Delgado, M. R., Stenger, V. A., and Fiez, J. A. (2004). Motivation-dependent responses in the human caudate nucleus. *Cereb. Cortex* 14, 1022–1030. doi: 10.1093/cercor/bhh062

Desai, R., Liebenthal, E., Waldron, E., and Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *J. Cogn. Neurosci.* 20, 1174–1188. doi: 10.1162/jocn.2008.20081

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1245–1255. doi: 10.1098/rstb.1998.0280

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* 22, 567–631. doi: 10.1146/annurev.neuro.22.1.567

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* 12, 961–974.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr. Opin. Neurobiol.* 10, 732–739. doi: 10.1016/S0959-4388(00)00153-7

Duncan, J., Martens, S., and Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature* 387, 808–810. doi: 10.1038/42947

Elliott, R., Frith, C. D., and Dolan, R. J. (1997). Differential neural response to positive and negative feedback in planning and guessing tasks. *Neuropsychologia* 35, 1395–1404.

Elliott, R., Newman, J. L., Longe, O. A., and Deakin, J. F. W. (2004). Instrumental responding for rewards is associated with enhanced neuronal response in subcortical reward systems. *Neuroimage* 21, 984–990. doi: 10.1016/j.neuroimage.2003.10.010

Filoteo, J. V., Maddox, W. T., and Davis, J. D. (2001). A possible role of the striatum in linear and nonlinear category learning: evidence from patients with Huntington's disease. *Behav. Neurosci.* 115, 786–798. doi: 10.1037/0735-7044.115.4.786

Flaherty, A. W., and Graybiel, A. M. (1993). Two input systems for body representations in the primate striatal matrix: experimental evidence in the squirrel monkey. *J. Neurosci.* 13, 1120–1137.

Flaherty, A. W., and Graybiel, A. M. (1994). Input-output organization of the sensorimotor striatum in the squirrel monkey. *J. Neurosci.* 14, 599–610.

Flege, J. E. (1995). "Second language speech learning theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed W. Strange (Timonium, MD: York Press), 233–277.

Folstein, J. R., Gauthier, I., and Palmeri, T. J. (2012). How category learning affects object representations: not all morphspaces stretch alike. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 807–820. doi: 10.1037/a0025836

Folstein, J. R., Palmeri, T. J., and Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cereb. Cortex* 23, 814–823. doi: 10.1093/cercor/bhs067

Foote, S. L., and Morrison, J. H. (1987). Extrathalamic modulation of cortical function. *Annu. Rev. Neurosci.* 10, 67–95. doi: 10.1146/annurev.ne.10.030187.000435

Francis, A. L., Baldwin, K., and Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Percept. Psychophys.* 62, 1668–1680. doi: 10.3758/BF03212164

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *J. Phon.* 36, 268–294. doi: 10.1016/j.wocn.2007.06.005

Francis, A. L., and Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 349–366. doi: 10.1037//0096-1523.28.2.349

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316. doi: 10.1126/science.291.5502.312

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.* 23, 5235–5246.

Fuster, J. M., Bauer, R. H., and Jervey, J. P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Res.* 330, 299–307. doi: 10.1016/0006-8993(85)90689-4

Gaffan, D., and Eacott, M. J. (1995). Visual learning for an auditory secondary reinforcer by macaques is intact after uncinate fascicle section: indirect evidence for the involvement of the corpus striatum. *Eur. J. Neurosci.* 7, 1866–1871. doi: 10.1111/j.1460-9568.1995.tb00707.x

Geiser, E., Notter, M., and Gabrieli, J. D. E. (2012). A corticostriatal neural system enhances auditory perception through temporal context processing. *J. Neurosci.* 32, 6177–6182. doi: 10.1523/JNEUROSCI.5153-11.2012

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

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *J. Exp. Psychol. Gen.* 123, 178–200. doi: 10.1037/0096-3445.123.2.178

Golestani, N., and Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage* 21, 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Gordon, P. C., Keyes, L., and Yung, Y. F. (2001). Ability in perceiving nonnative contrasts: performance on natural and synthetic speech stimuli. *Percept. Psychophys.* 63, 746–758. doi: 10.3758/BF03194435

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R." *Neuropsychologia* 9, 317–323. doi: 10.1016/0028-3932(71)90027-3

Goudbeek, M., Cutler, A., and Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Commun.* 50, 109–125. doi: 10.1016/j.specom.2007.07.003

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, 509–516. doi: 10.1177/0165025406071914

Guediche, S., Blumstein, S. E., Fiez, J. A., and Holt, L. L. (2014). Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Front. Syst. Neurosci.* 7:126. doi: 10.3389/fnsys.2013.00126

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychol. Rev.* 102, 594–621. doi: 10.1037/0033-295X.102.3.594

Guenther, F. H., and Ghosh, S. S. (2003). "A model of cortical and cerebellar function in speech," in *Proceedings of the XVth International Congress of Phonetic Sciences* (Barcelona), 169–173.

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., and Tourville, J. A. (2004). Representation of sound categories in auditory cortical maps. *J. Speech Lang. Hear. Res.* 47, 46–57. doi: 10.1044/1092-4388(2004/005)

Gureckis, T. M., and Goldstone, R. L. (2008). "The effect of the internal structure of categories on perception," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (Austin, TX), 1876–1881.

Gutnisky, D. A., Hansen, B. J., Iliescu, B. F., and Dragoi, V. (2009). Attention alters visual plasticity during exposure-based learning. *Curr. Biol.* 19, 555–560. doi: 10.1016/j.cub.2009.01.063

Haber, S. N., and Fudge, J. L. (1997). The primate substantia nigra and VTA: integrative circuitry and function. *Crit. Rev. Neurobiol.* 11, 323–342. doi: 10.1615/CritRevNeurobiol.v11.i4.40

Han, S., Huettel, S. A., Raposo, A., Adcock, R. A., and Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: recovery or goal attainment? *J. Neurosci.* 30, 4767–4775. doi: 10.1523/JNEUROSCI.3077-09.2010

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., et al. (2004). A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J. Neurosci.* 24, 1660–1665. doi: 10.1523/JNEUROSCI.3417-03.2004

Heald, S. L. M., and Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Front. Syst. Neurosci.* 8:35. doi: 10.3389/fnsys.2014.00035

Hedreen, J. C., and DeLong, M. R. (1991). Organization of striatopallidal, striatonigral, and nigrostriatal projections in the macaque. *J. Comp. Neurol.* 304, 569–595. doi: 10.1002/cne.903040406

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hikosaka, O., Sakamoto, M., and Usui, S. (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J. Neurophysiol.* 61, 814–832.

Hochstenbach, J., Spaendonck, K. P. V., Cools, A. R., Horstink, M. W., and Mulder, T. (1998). Cognitive deficits following stroke in the basal ganglia. *Clin. Rehabil.* 12, 514–520. doi: 10.1191/026921598666870672

Hökfelt, T., Johansson, O., Fuxe, K., Goldstein, M., and Park, D. (1977). Immunohistochemical studies on the localization and distribution of monoamine neuron systems in the rat brain II. Tyrosine hydroxylase in the telencephalon. *Med. Biol.* 55, 21–40.

Hökfelt, T., Ljungdahl, A., Fuxe, K., and Johansson, O. (1974). Dopamine nerve terminals in the rat limbic cortex: aspects of the dopamine hypothesis of schizophrenia. *Science* 184, 177–179. doi: 10.1126/science.184.4133.177

Hollerman, J. R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* 1, 304–309. doi: 10.1038/1124

Hollich, G., Newman, R. S., and Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Dev.* 76, 598–613. doi: 10.1111/j.1467-8624.2005.00866.x

Holt, L. L., and Lotto, A. J. (2006). Cue weighting in auditory categorization: implications for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059. doi: 10.1121/1.2188377

Holt, L. L., and Lotto, A. J. (2010). Speech perception as categorization. *Atten. Percept. Psychophys.* 72, 1218–1227. doi: 10.3758/APP.72.5.1218

Houk, J. C., and Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cereb. Cortex* 5, 95–110.

Idemaru, K., and Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1939–1956. doi: 10.1037/a0025641

Idemaru, K., Holt, L. L., and Seltman, H. (2012). Individual differences in cue weights are stable across time: the case of Japanese stop lengths. *J. Acoust. Soc. Am.* 132, 3950–3964. doi: 10.1121/1.4765076

Ingvalson, E. M., Holt, L. L., and McClelland, J. L. (2011). Can native Japanese listeners learn to differentiate /r–l/ on the basis of F3 onset frequency? *Biling. Lang. Cogn.* 15, 255–274. doi: 10.1017/S1366728911000447

Iverson, P., Hazan, V., and Bannister, K. (2005). Phonetic training with acoustic cue manipulations: a comparison of methods for teaching English /r/-/l/ to Japanese adults. *J. Acoust. Soc. Am.* 118, 3267. doi: 10.1121/1.2062307

Iverson, P., Kuhl, P. K., Akahane-yamada, R., and Diesch, E. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, 47–57. doi: 10.1016/S0010-0277(02)00198-1

Jacobs, D. H., Shuren, J., and Heilman, K. M. (1995). Impaired perception of facial identity and facial affect in Huntington's disease. *Neurology* 45, 1217–1218. doi: 10.1212/WNL.45.6.1217

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., Vanmeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. *Neuron* 53, 891–903. doi: 10.1016/j.neuron.2007.02.015

Joel, D., and Weiner, I. (1994). The organization of the basal ganglia-thalamocortical circuits: open interconnected rather than closed segregated. *Neuroscience* 63, 363–379. doi: 10.1016/0306-4522(94)90536-3

Kähkönen, S., Ahveninen, J., Jääskeläinen, I. P., Kaakkola, S., Näätänen, R., Huttunen, J., et al. (2001). Effects of haloperidol on selective attention: a combined whole-head MEG and high-resolution EEG study. *Neuropsychopharmacology* 25, 498–504. doi: 10.1016/S0893-133X(01)00255-X

Kemp, J. M., and Powell, T. P. (1971). The connexions of the striatum and globus pallidus: synthesis and speculation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 262, 441–457. doi: 10.1098/rstb.1971.0106

Kerr, J. N., and Wickens, J. R. (2001). Dopamine D-1/D-5 receptor activation is required for long-term potentiation in the rat neostriatum *in vitro*. *J. Neurophysiol.* 85, 117–124.

Kim, H. F., and Hikosaka, O. (2013). Distinct Basal Ganglia circuits controlling behaviors guided by flexible and stable values. *Neuron* 79, 1001–1010. doi: 10.1016/j.neuron.2013.06.044

Kim, J. N., and Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nat. Neurosci.* 2, 176–185. doi: 10.1038/5739

Koepp, M. J., Gunn, R. N., Lawrence, A. D., Cunningham, V. J., Dagher, A., Jones, T., et al. (1998). Evidence for striatal dopamine release during a video game. *Nature* 393, 266–268. doi: 10.1038/30498

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

Kuhl, P. K. (2003). Human speech and birdsong: communication and the social brain. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9645–9646. doi: 10.1073/pnas.1733998100

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533

Kuhl, P. K. (2007). Is speech learning "gated" by the social brain? *Dev. Sci.* 10, 110–120. doi: 10.1111/j.1467-7687.2007.00572.x

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. Sci.* 9, F13–F21. doi: 10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9096–9101. doi: 10.1073/pnas.1532872100

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255, 606–608. doi: 10.1126/science.1736364

Lawrence, A. D., Sahakian, B. J., and Robbins, T. W. (1998). Cognitive functions and corticostriatal circuits: insights from Huntington's disease. *Trends Cogn. Sci.* 2, 379–388. doi: 10.1016/S1364-6613(98)01231-5

Lee, Y.-S., Turkeltaub, P., Granger, R., and Raizada, R. D. S. (2012). Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. *J. Neurosci.* 32, 3942–3948. doi: 10.1523/JNEUROSCI.3814-11.2012

Leech, R., Holt, L. L., Devlin, J. T., and Dick, F. (2009). Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. *J. Neurosci.* 29, 5234–5239. doi: 10.1523/JNEUROSCI.5758-08.2009

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., and Formisano, E. (2012). Learning of new sound categories shapes neural response patterns in human auditory cortex. *J. Neurosci.* 32, 13273–13280. doi: 10.1523/JNEUROSCI.0584-12.2012

Liberman, A. M. (1996). *Speech: A Special Code.* Cambridge, MA: MIT Press.

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

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cereb. Cortex* 15, 1621–1631. doi: 10.1093/cercor/bhi040

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970. doi: 10.1093/cercor/bhq045

Lim, S.-J., and Holt, L. L. (2011). Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. *Cogn. Sci.* 35, 1390–1405. doi: 10.1111/j.1551-6709.2011.01192.x

Lim, S.-J., Holt, L. L., and Fiez, J. A. (2013). "Context-dependent modulation of striatal systems during incidental auditory category learning," in *Poster Presented at the Annual Meeting of the Society for Neuroscience* (San Diego, CA).

Lindvall, O., Björklund, A., Moore, R. Y., and Stenevi, U. (1974). Mesencephalic dopamine neurons projecting to neocortex. *Brain Res.* 81, 325–331. doi: 10.1016/0006-8993(74)90947-0

Lisker, L. (1986). "Voicing" in english: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang. Speech* 29, 3–11.

Liu, R., and Holt, L. L. (2011). Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. *J. Cogn. Neurosci.* 23, 1–16. doi: 10.1162/jocn.2009.21392

Lively, S. E., Logan, J. S., and Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: the role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* 94(3 pt 1), 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., and Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *J. Acoust. Soc. Am.* 96, 2076–2087.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (1991). Training Japanese listeners to identify English / r / and / 1 /: a first report for publication. *J. Acoust. Soc. Am.* 89, 874–886. doi: 10.1121/1.1894649

Lopez-Paniagua, D., and Seger, C. A. (2011). Interactions within and between corticostriatal loops during component processes of category learning. *J. Cogn. Neurosci.* 23, 3068–3083. doi: 10.1162/jocn_a_00008

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

Lotto, A. J., Sato, M., and Diehl, R. L. (2004). "Mapping the task for the second language learner: the case of Japanese acquisition of /r/ and /l/," in *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, eds J. Slifka, S. Manuel, and M. Matthies (Cambridge, MA: MIT), 181–186.

Lynd-Balta, E., and Haber, S. N. (1994). The organization of midbrain projections to the striatum in the primate: sensorimotor-related striatum versus ventral striatum. *Neuroscience* 59, 625–640. doi: 10.1016/0306-4522(94)90182-1

Maddox, W. T., Ashby, F. G., and Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 650–662. doi: 10.1037/0278-7393.29.4.650

Maddox, W. T., Bohil, C. J., and Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychon. Bull. Rev.* 11, 945–952. doi: 10.3758/BF03196726

Maddox, W. T., Love, B. C., Glass, B. D., and Filoteo, J. V. (2008). When more is less: feedback effects in perceptual category learning. *Cognition* 108, 578–589. doi: 10.1016/j.cognition.2008.03.010

Malachi, R., and Graybiel, A. M. (1986). Mosaic architecture of the somatic sensory-recipient sector of the cat's striatum. *J. Neurosci.* 6, 3436–3458.

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, B101–B11. doi: 10.1016/S0010-0277(01)00157-3

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., and McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cogn. Affect. Behav. Neurosci.* 2, 89–108. doi: 10.3758/CABN.2.2.89

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McClelland, J. L., Fiez, J. A., and McCandliss, B. D. (2002). Teaching the /r/-/l/ discrimination to Japanese adults: behavioral and neural aspects. *Physiol. Behav.* 77, 657–662. doi: 10.1016/S0031-9384(02)00916-2

McClelland, J. L., Thomas, A. G., McCandliss, B. D., and Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. *Prog. Brain Res.* 121, 75–80. doi: 10.1016/S0079-6123(08)63068-X

McClure, S. M., Berns, G. S., and Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38, 339–346. doi: 10.1016/S0896-6273(03)00154-5

McMurray, B., Aslin, R. N., and Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Dev. Sci.* 12, 369–378. doi: 10.1111/j.1467-7687.2009.00822.x

McNamee, D., Rangel, A., and O'Doherty, J. P. (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nat. Neurosci.* 16, 479–485. doi: 10.1038/nn.3337

Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9014–9019. doi: 10.1073/pnas.1105040108

Mehler, J., Jusczyk, P., Lamsertz, G., and French, F. (1988). A precursor of language acquisition in young infants. *Cognition* 29, 143–178. doi: 10.1016/0010-0277(88)90035-2

Middleton, F. A., and Strick, P. L. (1996). The temporal lobe is a target of output from the basal ganglia. *Proc. Natl. Acad. Sci. U.S.A.* 93, 8683–8687. doi: 10.1073/pnas.93.16.8683

Middleton, F. A., and Strick, P. L. (2000). Basal ganglia output and cognition: evidence from anatomical, behavioral, and clinical studies. *Brain Cogn.* 42, 183–200. doi: 10.1006/brcg.1999.1099

Miller, B. T., Vytlacil, J., Fegen, D., Pradhan, S., and D'Esposito, M. (2011). The prefrontal cortex modulates category selectivity in human extrastriate cortex. *J. Cogn. Neurosci.* 23, 1–10. doi: 10.1162/jocn.2010.21516

Miller, E. K., and Buschman, T. (2008). "Rules through recursion: how interactions between the frontal cortex and basal ganglia may build abstract, complex rules from concrete, simple ones," in *Neuroscience of Rule-Guided Behavior*, eds S. A. Bunge and J. D. Wallis (New York, NY: Oxford University Press), 419–440.

Mishkin, M. (1982). A memory system in the monkey. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 298, 85–95. doi: 10.1098/rstb.1982.0074

Mishkin, M., Malamut, B., and Bachevalier, J. (1984). "Memories and habits: two neural systems," in *Neurobiology of Learning and Memory*, eds G. Lynch, J. L. McGaugh, and N. M. Weinberger (New York, NY: The Guilford Press), 65–77.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (1975). An effect of linguistic experience: the discrimination of /r/ and /l/ by native speakers of Japanese and English. *Percept. Psychophys.* 18, 331–340. doi: 10.3758/BF03211209

Moon, C., Cooper, R. P., and Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behav. Dev.* 16, 495–500. doi: 10.1016/0163-6383(93)80007-U

Moriizumi, T., and Hattori, T. (1992). Separate neuronal populations of the rat globus pallidus projecting to the subthalamic nucleus, auditory cortex and pedunculopontine tegmental area. *Neuroscience* 46, 701–710. doi: 10.1016/0306-4522(92)90156-V

Moriizumi, T., Nakamura, Y., Tokuno, H., Kitao, Y., and Kudo, M. (1988). Topographic projections from the basal ganglia to the nucleus tegmenti pedunculopontinus pars compacta of the cat with special reference to pallidal projections. *Exp. Brain Res.* 71, 298–306. doi: 10.1007/BF00247490

Muhammad, R., Wallis, J. D., and Miller, E. K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, inferior temporal cortex, and striatum. *J. Cogn. Neurosci.* 18, 974–989. doi: 10.1162/jocn.2006.18.6.974

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *J. Acoust. Soc. Am.* 85, 365–378. doi: 10.1121/1.397688

Nauta, H. J., Pritz, M. B., and Lasek, R. J. (1974). Afferents to the rat caudoputamen studied with horseradish peroxidase. An evaluation of a retrograde neuroanatomical research method. *Brain Res.* 67, 219–238.

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., and Parrish, T. B. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cereb. Cortex* 17, 37–43. doi: 10.1093/cercor/bhj122

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304, 452–454. doi: 10.1126/science.1094285

Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., and Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *J. Neurosci.* 26, 13025–13036. doi: 10.1523/JNEUROSCI.2481-06.2006

Packard, M. G., Hirsh, R., and White, N. M. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. *J. Neurosci.* 9, 1465–1472.

Packard, M. G., and McGaugh, J. L. (1992). Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: further evidence for multiple memory systems. *Behav. Neurosci.* 106, 439–446. doi: 10.1037/0735-7044.106.3.439

Palmeri, T. J., and Gauthier, I. (2004). Visual object understanding. *Nat. Rev. Neurosci.* 5, 291–303. doi: 10.1038/nrn1364

Parent, A., Boucher, R., and O'Reilly-Fromentin, J. (1981). Acetylcholinesterase-containing neurons in cat pallidal complex: morpho- logical characteristics and projection towards the neocortex. *Brain Res.* 230, 356–361.

Parent, A., and Hazrati, L. N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res. Brain Res. Rev.* 20, 91–127. doi: 10.1016/0165-0173(94)00007-C

Pasupathy, A., and Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* 433, 873–876. doi: 10.1038/nature03287

Percheron, G., Francois, C., Yelnik, J., Fenelon, G., and Talbi, B. (1994). "The basal ganglia related systems of primates: definition, description and informational analysis," in *The Basal Ganglia IV*, eds G. Percheron, G. M. McKenzie, and J. Feger (New York, NY: Plenum Press), 3–20.

Petrides, M. (1985). Deficits in non-spatial conditional associative learning after periarcuate lesions in the monkey. *Behav. Brain Res.* 16, 95–101. doi: 10.1016/0166-4328(85)90085-3

Pisoni, D. B. (1992). "Some comments on invariance, variability, and perceptual normalization in speech perception," in *Proceedings of the International Conference on Spoken Language Processing* (Banff, AB), 587–590.

Posner, M. I., and Keele, S. W. (1968). On the genesis of abstract ideas. *J. Exp. Psychol.* 77, 353–363. doi: 10.1037/h0025953

Reynolds, J. N. J., and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw.* 15, 507–521. doi: 10.1016/S0893-6080(02)00045-X

Robbins, T. W., and Everitt, B. J. (1992). Functions of dopamine in the dorsal and ventral striatum. *Semin. Neurosci.* 4, 119–127. doi: 10.1016/1044-5765(92)90010-Y

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition* 70, 27–52. doi: 10.1016/S0010-0277(98)00075-4

Saint-Cyr, J. A. (2003). Frontal-striatal circuit functions: context, sequence, and consequence. *J. Int. Neuropsychol. Soc.* 9, 103–127. doi: 10.1017/S1355617703910125

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.

Schultz, W. (1999). The reward signal of midbrain dopamine neurons. *News Physiol. Sci.* 14, 249–255.

Schultz, W. (2000). Multiple reward signals in the brain. *Nat. Rev. Neurosci.* 1, 199–207. doi: 10.1038/35044563

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron* 36, 241–263. doi: 10.1016/S0896-6273(02)00967-4

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.* 13, 900–913.

Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.* 12, 4595–4610.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci. Biobehav. Rev.* 32, 265–278. doi: 10.1016/j.neubiorev.2007.07.010

Seger, C. A. (2013). The visual corticostriatal loop through the tail of the caudate: circuitry and function. *Front. Syst. Neurosci.* 7:104. doi: 10.3389/fnsys.2013.00104

Seger, C. A., and Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *J. Neurosci.* 25, 2941–2951. doi: 10.1523/JNEUROSCI.3401-04.2005

Seger, C. A., and Miller, E. K. (2010). Category learning in the brain. *Annu. Rev. Neurosci.* 33, 203–219. doi: 10.1146/annurev.neuro.051508.135546

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., and Anderson, C. W. (2010). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and Granger causality modeling. *Neuroimage* 50, 644–656. doi: 10.1016/j.neuroimage.2009.11.083

Seitz, A. R., Kim, D., and Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron* 61, 700–707. doi: 10.1016/j.neuron.2009.01.016

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., et al. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition* 115, 435–443. doi: 10.1016/j.cognition.2010.03.004

Seitz, A. R., and Watanabe, T. (2003). Is subliminal learning really passive? *Nature* 422, 36. doi: 10.1038/422036a

Seitz, A. R., and Watanabe, T. (2005). A unified model for perceptual learning. *Trends Cogn. Sci.* 9, 329–334. doi: 10.1016/j.tics.2005.05.010

Seitz, A. R., and Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Res.* 49, 2604–2610. doi: 10.1016/j.visres.2009.08.003

Selemon, L. D., and Goldman-Rakic, P. S. (1985). Longitudinal topography and lnterdigitation projections in the rhesus monkey. *J. Neurosci.* 5, 776–794.

Selemon, L. D., and Goldman-Rakic, P. S. (1990). Topographic intermingling of striatonigral and striatopallidal neurons in the rhesus monkey. *J. Comp. Neurol.* 297, 359–376. doi: 10.1002/cne.902970304

Sigala, N., and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415, 318–320. doi: 10.1038/415318a

Silkis, I. (2007). A hypothetical role of cortico-basal ganglia-thalamocortical loops in visual processing. *Biosystems* 89, 227–235. doi: 10.1016/j.biosystems.2006.04.020

Silkis, I. (2008). "Dopamine-dependent synaptic plasticity in the cortico-basal ganglia-thalamocortical loops as mechanism of visual attention," in *Synaptic Plasticity: New Research*, Vol. 7, eds E. T. F. Kaiser and F. J. Peters (New York, NY: Nova Science Publishers), 355–371.

Simon, H., Le Moal, M., and Calas, A. (1979). Efferents and afferents of the ventral tegmental-A10 region studied after local injection of [3H]leucine and horseradish peroxidase. *Brain Res.* 178, 17–40. doi: 10.1016/0006-8993(79)90085-4

Skinner, J. E., and Yingling, C. D. (1976). Regulation of slow potential shifts in nucleus reticularis thalami by the mesencephalic reticular formation and the frontal granular cortex. *Electroencephalogr. Clin. Neurophysiol.* 40, 288–296. doi: 10.1016/0013-4694(76)90152-8

Sutton, R. S., and Barto, A. G. (1998). Reinforcement learning: an introduction. *IEEE Trans. Neural Netw.* 9, 1054. doi: 10.1109/TNN.1998.712192

Swanson, L. W. (1982). The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Res. Bull.* 9, 321–353. doi: 10.1016/0361-9230(82)90145-9

Szabo, J. (1979). Strionigral and nigrostriatal connections. Anatomical studies. *Appl. Neurophysiol.* 42, 9–12.

Teinonen, T., Aslin, R. N., Alku, P., and Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition* 108, 850–855. doi: 10.1016/j.cognition.2008.05.009

Thierry, A. M., Blanc, G., Sobel, A., Stinus, L., and Golwinski, J. (1973). Dopaminergic terminals in the rat cortex. *Science* 182, 499–501. doi: 10.1126/science.182.4111.499

Thiessen, E. D. (2010). Effects of visual information on adults' and infants' auditory statistical learning. *Cogn. Sci.* 34, 1093–1106. doi: 10.1111/j.1551-6709.2010.01118.x

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies.* New York, NY: Macmillan.

Toro, J. M., Sinnett, S., and Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition* 97, B25–B34. doi: 10.1016/j.cognition.2005.01.006

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Tremblay, L., Hollerman, J. R., and Schultz, W. (1998). Modifications of reward expectation-related neuronal activity during learning in primate striatum. *J. Neurophysiol.* 80, 964–977.

Tricomi, E., Delgado, M. R., and Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron* 41, 281–292. doi: 10.1016/S0896-6273(03)00848-1

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., and Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *J. Cogn. Neurosci.* 18, 1029–1043. doi: 10.1162/jocn.2006.18.6.1029

Tricomi, E., and Fiez, J. A. (2008). Feedback signals in the caudate reflect goal achievement on a declarative memory task. *Neuroimage* 41, 1154–1167. doi: 10.1016/j.neuroimage.2008.02.066

Tricomi, E., and Fiez, J. A. (2012). Information content and reward processing in the human striatum during performance of a declarative memory task. *Cogn. Affect. Behav. Neurosci.* 12, 361–372. doi: 10.3758/s13415-011-0077-3

Tsushima, Y., Seitz, A. R., and Watanabe, T. (2008). Task-irrelevant learning occurs only when the irrelevant feature is weak. *Curr. Biol.* 18, R516–R517. doi: 10.1016/j.cub.2008.04.029

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., et al. (1997). A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *J. Cogn. Neurosci.* 9, 266–276. doi: 10.1162/jocn.1997.9.2.266

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems," in *Analysis of Visual Behavior*, eds D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Cambridge, MA: MIT Press), 549–586.

Vallabha, G. K., and McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: consequences of learned Hebbian attractors in topographic maps. *Cogn. Affect. Behav. Neurosci.* 7, 53–73. doi: 10.3758/CABN.7.1.53

van der Linden, M., Wegman, J., and Fernández, G. (2014). Task- and experience-dependent cortical selectivity to features informative for categorization. *J. Cogn. Neurosci.* 26, 319–333. doi: 10.1162/jocn_a_00484

Van Hoesen, G. W., Yeterian, E. H., and Lavizzo-Mourey, R. (1981). Widespread corticostriate projections from temporal cortex of the rhesus monkey. *J. Comp. Neurol.* 199, 205–219. doi: 10.1002/cne.901990205

van Schouwenburg, M. R., den Ouden, H. E. M., and Cools, R. (2010). The human basal ganglia modulate frontal-posterior connectivity during attention shifting. *J. Neurosci.* 30, 9910–9918. doi: 10.1523/JNEUROSCI.1111-10.2010

Vlahou, E. L., Protopapas, A., and Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *J. Exp. Psychol. Gen.* 141, 363–381. doi: 10.1037/a0025014

Wallis, J. D., Anderson, K. C., and Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature* 411, 953–956. doi: 10.1038/35082081

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *J. Acoust. Soc. Am.* 106, 3649–3658. doi: 10.1121/1.428217

Wade, T., and Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *J. Acoust. Soc. Am.* 118, 2618. doi: 10.1121/1.2011156

Webster, K. E. (1961). Cortico-striate interrelations in the albino rat. *J. Anat.* 95, 532–544.

Werker, J. F., and Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Percept. Psychophys.* 37, 35–44. doi: 10.3758/BF03207136

Werker, J. F., and Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* 7, 49–63. doi: 10.1016/S0163-6383(84)80022-3

Wickens, J. R. (1997). Basal ganglia: structure and computations. *Netw. Comput. Neural Syst.* 8, 77–109. doi: 10.1088/0954-898X/8/4/001

Wickens, J. R., Begg, A. J., and Arbuthnott, G. W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex *in vitro*. *Neuroscience*, 70, 1–5. doi: 10.1016/0306-4522(95)00436-M

Wickens, J. R., Reynolds, J. N. J., and Hyland, B. I. (2003). Neural mechanisms of reward-related motor learning. *Curr. Opin. Neurobiol.* 13, 685–690. doi: 10.1016/j.conb.2003.10.013

Wilson, C. J. (1995). "The contribution of cortical neurons to the firing pattern of striatal spiny neurons," in *Models of Information Processing in the Basal Ganglia*, eds J. C. Houk, J. L. Davis, and D. G. Beiser (Cambridge, MA: Bradford), 29–50.

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

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702. doi: 10.1038/nn1263

Wise, R. A., and Rompre, P. P. (1989). Brain and dopamine reward. *Annu. Rev. Psychol.* 40, 191–225. doi: 10.1146/annurev.psych.40.1.191

Yamamoto, S., Kim, H. F., and Hikosaka, O. (2013). Reward value-contingent changes of visual responses in the primate caudate tail associated with a visuomotor skill. *J. Neurosci.* 33, 11227–11238. doi: 10.1523/JNEUROSCI.0318-13.2013

Yeterian, E. H., and Pandya, D. N. (1998). Corticostriatal connections of the superior temporal region in rhesus monkeys. *J. Comp. Neurol.* 399, 384–402.

Yeung, H. H., and Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition* 113, 234–243. doi: 10.1016/j.cognition.2009.08.010

Zheng, T., and Wilson, C. J. (2002). Corticostriatal combinatorics: the implications of corticostriatal axonal arborizations. *J. Neurophysiol.* 87, 1007–1017.

Znamenskiy, P., and Zador, A. M. (2013). Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* 497, 482–485. doi: 10.1038/nature12077

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: 23 April 2014; accepted: 13 July 2014; published online: 01 August 2014.
*Citation: Lim S-J, Fiez JA and Holt LL (2014) How may the basal ganglia contribute to auditory categorization and speech perception? Front. Neurosci. **8**:230. doi: 10.3389/fnins.2014.00230*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 18 March 2013
doi: 10.3389/fnins.2013.00034

Auditory perceptual learning for speech perception can be enhanced by audiovisual training

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

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

Edited by:
Claude Alain, Rotman Research Institute, Canada

Reviewed by:
Mark T. Wallace, Vanderbilt University, USA
Iiro P. Jääskeläinen, Aalto University, Finland

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

Speech perception under audiovisual (AV) conditions is well known to confer benefits to perception such as increased speed and accuracy. Here, we investigated how AV training might benefit or impede auditory perceptual learning of speech degraded by vocoding. In Experiments 1 and 3, participants learned paired associations between vocoded spoken nonsense words and nonsense pictures. In Experiment 1, paired-associates (PA) AV training of one group of participants was compared with audio-only (AO) training of another group. When tested under AO conditions, the AV-trained group was significantly more accurate than the AO-trained group. In addition, pre- and post-training AO forced-choice consonant identification with untrained nonsense words showed that AV-trained participants had learned significantly more than AO participants. The pattern of results pointed to their having learned at the level of the auditory phonetic features of the vocoded stimuli. Experiment 2, a no-training control with testing and re-testing on the AO consonant identification, showed that the controls were as accurate as the AO-trained participants in Experiment 1 but less accurate than the AV-trained participants. In Experiment 3, PA training alternated AV and AO conditions on a list-by-list basis within participants, and training was to criterion (92% correct). PA training with AO stimuli was reliably more effective than training with AV stimuli. We explain these discrepant results in terms of the so-called "reverse hierarchy theory" of perceptual learning and in terms of the diverse multisensory and unisensory processing resources available to speech perception. We propose that early AV speech integration can potentially impede auditory perceptual learning; but visual top-down access to relevant auditory features can promote auditory perceptual learning.

Keywords: audiovisual speech processing, audiovisual speech perception, perceptual learning, reverse hierarchy theory, auditory perception, visual speech perception, multisensory processing, plasticity and learning

INTRODUCTION

In addition to the classically defined, high-level multisensory cortical association areas such as the superior temporal sulcus (Calvert et al., 2000; Beauchamp et al., 2004; Miller and D'Esposito, 2005; Nath and Beauchamp, 2012), multisensory processing sites have been identified at lower levels, such as primary or secondary cortical areas and the major thalamic relay nuclei (for reviews, see Foxe and Schroeder, 2005; Driver and Noesselt, 2008; Falchier et al., 2012; Kayser et al., 2012). For example, monkey studies have found visual neuronal inputs to primary auditory cortex and to the caudal auditory belt cortex (Schroeder and Foxe, 2002; Ghazanfar et al., 2005; Kayser et al., 2009). Evidence is also available for auditory neuronal inputs to primary visual cortex (Falchier et al., 2001, 2012). Extensive multisensory connectivity has led to the suggestion that all cortical operations are potentially multisensory (Ghazanfar and Schroeder, 2006).

There is no doubt that speech perception makes use of diverse multisensory cortical processing resources (Sams et al., 1991; Calvert et al., 2000; Möttönen et al., 2002; Miller and D'Esposito, 2005; Saint-Amour et al., 2007; Skipper et al., 2007; Bernstein et al., 2008a,b; Nath and Beauchamp, 2011, 2012), and that visual speech stimuli integrate with auditory stimuli under a wide range of listening conditions and for a wide range of functions. For example, when auditory speech stimuli are degraded, being able to see the talker typically leads to improved perceptual accuracy (e.g., Sumby and Pollack, 1954; MacLeod and Summerfield, 1987; Iverson et al., 1998; Ross et al., 2007; Ma et al., 2009). But even when the auditory stimuli are not degraded, visual speech stimuli can affect speech perception and comprehension. Comprehension of difficult verbal materials can be easier under audiovisual (AV) conditions (Reisberg et al., 1987); Perception in a second language can be more accurate with AV stimuli than with auditory-only stimuli (Hazan et al., 2006); and Numerous demonstrations of the McGurk effect (McGurk and MacDonald, 1976) have shown that when auditory and visual speech consonants are mismatched, perceivers often hear a consonant that is different from either the auditory or visual stimulus *per se* (e.g., Green and Kuhl, 1989; Sekiyama and Tohkura, 1991; Jiang and Bernstein, 2011). The study reported here addressed how training with AV speech stimuli might affect auditory perceptual learning of a type of novel degraded acoustic speech stimulus. At issue was how multisensory resources are deployed in the context of unisensory perceptual learning.

This study focused on learning to perceive degraded acoustic speech. The spoken nonsense words that were used as stimuli were transformed by passing them through a vocoder, a signal-processor that systematically degrades the speech (Iverson et al., 1998; Scott et al., 2000) and typically requires experience or training to achieve improved levels of perceptual accuracy (e.g., Davis et al., 2005; Scott et al., 2006; Hervais-Adelman et al., 2011). The vocoder here transformed fine-grained acoustic spectral cues, including vocal tract resonance changes that are cues to phoneme (consonants and vowels) distinctions, into coarse spectral cues by coding energy in 15 frequency bands as amplitudes of fixed-frequency sinusoids at the center frequency of each band (**Figure 1**). In addition, the normal speech spectrum, which falls off at approximately 6 dB per octave, was tilted so that amplitudes in vocoder bands were approximately equalized. **Figure 1** shows spectrograms of the syllables /bE/ and /fE/ (i.e., the vowel in "bet") for the natural recorded speech (**Figures 1A,C**) and the vocoded speech (**Figures 1B,D**). The vocoding highly reduces the available acoustic information, emphasizes the second speech formant (vocal tract resonance), known to be highly informative for speech perception (Liberman et al., 1967), and reduces or omits the first and third formants, which are also important.

We hypothesized that information in visual speech stimuli can provide top-down guidance for auditory perceptual learning (Ahissar and Hochstein, 1997; Kral and Eggermont, 2007; Ahissar et al., 2008) of the cues to phoneme perception in the vocoded acoustic signals. That is, in addition to integrating with auditory speech cues during perception, visual speech stimuli were hypothesized to be able to guide auditory perceptual learning, with the result that auditory-only perception is improved more following AV than following auditory-only training. Our rationale for this hypothesis about the benefits of visual speech is that certain visual speech features can be reliably available (Bernstein et al., 2000; Bernstein, 2012), and they are correlated in real time with auditory features (Yehia et al., 1998; Jiang et al., 2002; Jiang and Bernstein, 2011). Therefore, they could help to train novel or unfamiliar vocoded auditory speech features when they are available during training. For example, /f/ and /b/ are visually distinctive (Auer and Bernstein, 1997), but the distinction between vocoded /f/ and /b/, which is available in the novel acoustic signals (see **Figures 1B,D**), might not be discerned without training. Training with the AV stimuli could enhance auditory perceptual learning, because the visual features that are integrated during visual perceptual processing (Bernstein et al., 2011; Bernstein, 2012) could be used to guide top-down attention to the correlated auditory cues that discriminate /f/ from /b/. In contrast, training with auditory-only stimuli contributes no additional information for learning novel cues or features, beyond what can be gleaned from merely repeating the stimulus, and the perceiver might not learn to distinguish the critical novel cues. Alternatively, early integration of auditory and visual speech features could impede auditory perceptual learning, because perception would be successful without accessing the available auditory distinctions in the vocoded stimuli.

In the study reported here, we compared auditory perceptual learning based on training with AV versus audio-only (AO) speech stimuli. Because our hypothesis concerned perceptual learning of acoustic speech features, the experimental task had to preclude access to pre-existing lexical knowledge, a type of high-level representation, that could function like visual speech stimuli. Lexical knowledge itself can be a top-down source for auditory perceptual learning (Davis et al., 2005). Therefore, all of the stimuli in the study were spoken nonsense words. Auditory training was given in a paired-associates (PA) task. Participants learned paired associations between disyllabic spoken nonsense words and nonsense pictures. Training was under AV and/or AO conditions, and testing was exclusively under AO conditions. In addition to PA training and testing, a forced-choice identification paradigm was used to test auditory consonant identification before and after training, using stimuli that were not used in training. The consonant identification also served to test for generalization to new stimuli in a different perceptual task and to infer the level of auditory perceptual learning that was achieved. Our results show that AV training can significantly benefit auditory perceptual learning beyond AO training. But the details of the training protocol appear to be critically important to achieving benefit from visual stimuli, because AV training can also lead to poorer AO performance. In our General Discussion, we propose a model of how AV stimuli can guide auditory perceptual learning through top-down visual access to useful auditory distinctions; or how AV stimuli can impede auditory perceptual learning through early immediate integration of auditory and visual speech cues.

MATERIALS AND METHODS

EXPERIMENT 1 BETWEEN-PARTICIPANT TRAINING WITH FIXED NUMBERS OF TRAINING TRIALS

In Experiment 1, participants were assigned to either AV or AO PA training followed by AO testing. Training in the PA task used nonsense pictures and nonsense words of the form consonant-vowel-consonant-vowel-consonant (CVCVC), modeled on the phonotactics of disyllabic English words. The PA task emulated the learning of new vocabulary items. Thus, participants were required to learn at multiple levels, including the perceptual (novel acoustic transform and novel lexical word form) and the high-level associative (semantic association between word form and picture). Here, participants were tested on the number of paired associations they could demonstrate following training. If AV-trained participants were more successful during AO testing than AO-trained participants, who had achieved equivalent performance during training, then the implication would be that the AV-trained participants learned more about the auditory stimuli. Pre- and post-training forced-choice consonant identification was tested, using an untrained set of CVCVC nonsense words. The identification measures were the number of correctly identified consonants in the three positions of the nonsense words. If differential learning occurred across the position of the consonant in the word, then the implication would be that participants learned sub-phonemic auditory features, because acoustic phonetic signals differ across segment position in a word (Stevens, 1998).

Subjects

Individuals were screened for American English as a first language, normal or corrected-to-normal vision in each eye of 20/30 or better (using a Snellen chart), and hearing (25 dB HL or better in each ear for frequencies between 125 and 8 KHz, using an Audiometrics

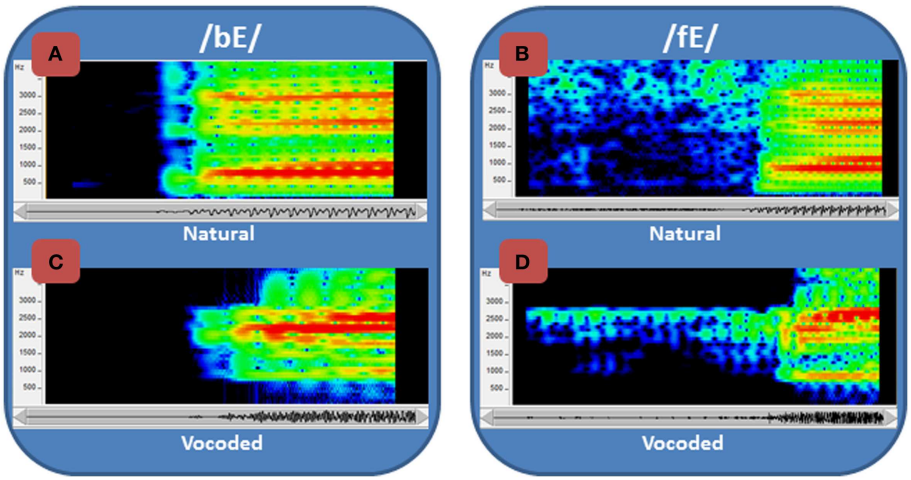


FIGURE 1 | Spectrograms of normal and vocoded speech. Spectrograms of speech show the concentrations of energy in the spectra over time. Two speech tokens, /bE/ and /fE/ (i.e., the vowel in "bet"), are shown in spectrograms of the natural **(A)** and **(B)** recorded speech and the vocoded **(C)** and **(D)** speech. The frequency range of the spectrograms is restricted to 4 kHz, because all of the energy from the vocoder is similarly limited. The amplitudes are represented as a heat map, with red the highest amplitude and dark blue the lowest. In addition to representing the speech as the sum of sinewaves at the center of each vocoder filter (see text), the vocoder also tilted the spectrum so that it did not roll off at approximately 6 dB/octave, which is natural to speech. Thus, the amplitudes of the frequencies vary across the natural and the vocoded speech, in addition to the frequency ranges and spectral detail.

GSI 16 audiometer with insert earphones). The experiment was carried out at two different locations, using the same equipment and procedures. At the House Research Institute (Los Angeles, CA, USA), 12 volunteers, ages 18–48 years (mean = 30 years), including six males, completed the experiment, and an additional five volunteers were asked to discontinue the experiment after they were mistakenly presented with non-distorted speech. At the George Washington University, 25 volunteers, ages 19–30 (mean = 22), including five males, completed the experiment, and an additional four dropped out due to lack of availability. In all, 18 participants completed AV training, and 19 completed AO training. They were paid $12 per hour of testing, plus any travel expenses incurred. Subjects gave written consent. Human subject participation was approved by either the St. Vincent's Hospital Institutional Review Board (Los Angeles, CA, USA) or by the George Washington University Institutional Review Board (Washington, DC, USA).

Stimuli

Speech. The spoken CVCVC nonsense words were modeled on English phonotactics (i.e., the sequential speech patterns in English). They were visually distinct for lipreading and visually unique from real English words (i.e., the words were designed to not be mistaken as real words, if they were lipread without accompanying audio). Thus, for example, the nonsense word *mucker* was not included in the set, because the visual stimulus could be mistaken for the real word *pucker*, inasmuch as the phonemes /p, m/ are visually highly similar (Auer and Bernstein, 1997).

The process of stimulus generation was as follows. Syllables with the structure CV-, -VCV-, and –VC were extracted from the 35,000-word phonemically transcribed PhLex database (Seitz et al., 1998). Based on empirically derived phonotactic probabilities, a Monte Carlo simulation was used to generate 30,000 CVCVC candidate nonsense words, which were then further processed. First, existing visual phoneme confusion data were used to model the confusability of the phonemes (Auer and Bernstein, 1997; Iverson et al., 1998). Then the candidate nonsense words were computationally processed, taking into account their visual confusability with real words and other nonsense words (Auer and Bernstein, 1997). Stimuli that would have been easily confused by vision were grouped into sets, and only one CVCVC word was chosen from each set, with the requirements that (1) the final set of nonsense words would include all the English phonemes, and (2) within each CVCVC, the five phonemes would be visually distinct to a lipreader (Auer and Bernstein, 1997). These constraints implied that within a list of nonsense words, visual information should be sufficient to differentiate among items.

The female talker whose data were used to model consonant and vowel confusability was the same talker used to produce the nonsense words. She was professionally videotaped uttering the final set of 260 CVCVC words.

Stimulus lists were constructed by first ordering stimuli by initial consonant and vowel, and then dividing the list on even- versus odd-numbered items to form two lists from which items were randomly selected. Two 49-item lists were selected for the pre- and post-training consonant identification task (**Table 1**; see **Table 2** for transcription key). Two six-item lists were selected from 12-item lists for pre- and post-training practice. Six lists of 12 items for PA training and six lists of six items as new items during PA testing were selected from the remaining available words (**Table 3**).

The acoustic speech stimuli were processed through a custom realtime hardware/software vocoder (Iverson et al., 1998). The vocoder detected speech energy in thirteen 120-Hz-bandwidth bandpass filters with center frequencies every 150 Hz from 825 Hz

Table 1 | Pre-test and post-test consonant identification lists in single-phoneme transcription format.

List 1		List 2	
banoz	pETat	batok	pod^n
biscg	ponRs	Bizxd	pUrIn
brcit	pUtIl	bRsxv	Ribcg
bulad	rid^t	bUnxl	rob^l
c@GRz	rot^k	C@pRk	s@naJ
ccrik	s@vxk	CctIG	SIGRt
cEmxl	sik^S	CEvxs	SInal
deman	Sivab	Dumxs	sRbik
duzxn	sRmaS	fRCxl	Sulak
fRsal	suZxm	gInxz	t@Cig
gIZxn	t@n^m	h@n^p	tEmaS
h@nus	tErin	Jcrat	Tib^n
jcrib	Tis^p	JEnap	Tufxl
jEris	Tukad	JozIG	v@sap
junxs	vEJUd	k@Cud	vEJxn
k@Taz	vob^n	Kcrit	vomit
kctas	vRbIG	m@DRz	vRlIs
m@JUd	Wcfxn	madRz	wct^m
makiz	wEJxk	Mckit	wEkab
mczin	wRk^l	mEros	wRlas
mezxl	Yizxk	nECUt	yiZxs
Net^m	yUbIg	Nobad	yUmEs
noluz	Yusap	p@Cik	yutIb
p@Tan	zobIG	paJUt	zoSxn
palIt		pEluz	

Words are transcribed, because English orthography does not map uniquely to English phonemes. ***Table 2*** *gives the phoneme transcription key. Lists 1 and 2 were randomly selected on a per-subject basis for use in pre-test and post-test (or test, re-test) consonant identification tasks. The practice list (JUkiz, zIJxl, dISus, JEroz, mivRd, DEkxs) was used before each test to ensure that participants understood the task.*

Table 2 | Transcription keys for nonsense word consonants and vowels.

Consonant sounds represented by lower case on keyboard		Consonant sounds represented by UPPER case on keyboard	
A			
Consonant	**Example**	**Consonant**	**Example**
b	(b)ut	C	su(ch)
d	goo(d)	D	(th)at
f	(f)ew	G	lo(ng)
g	(g)ood	J	lar(g)e
h	(h)is	S	(sh)e
k	(c)an	T	bo(th)
l	(l)ike	Z	u(s)ual
m	(m)ore		
n	(n)ew	consonants easily confused	
p	(p)ut	D	T
r	(r)oom	s	S
s	(s)ome	g	G
t	bu(t)	z	Z
v	gi(v)e	c	J
w	(w)ill	k	
y	(y)ou		
z	wa(s)		
B			
Vowel	**Example**	**Vowel**	**Example**
a	b(o)b	@	b(a)t
o	b(oa)t	E	b(e)t
i	b(ea)t	x	(a)bout
c	b(ou)ght	u	l(u)te
r	b(ir)d	I	b(i)t
u	b(oo)k	^	b(u)t

(A) Consonant transcription key. (B) Vowel transcription key. These transcription keys were used to assign a single orthographic symbol for each English consonant and vowel phoneme in the nonsense words listed in ***Tables 1*** *and* ***3****. The consonant transcription key was used to train and test participants to carry out forced-choice consonant identification.*

through 2625 Hz. Two additional filters were used to convey high frequencies. One was a bandpass filter centered at 3115 Hz with 350 Hz bandwidth and the other a highpass filter with 3565 Hz cutoff. The energy detected in each band was used to amplitude-modulate a fixed-frequency sinewave at the center frequency of that band (and at 3565 Hz in the case of the highpass filter). The sum of the 15 sinewaves comprised the vocoded acoustic signal. This acoustic transformation retained the gross spectral-temporal amplitude information in the waveform while eliminating finer distinctions such as fundamental frequency variations and eliminating the natural spectral tilt of the vocal tract resonances. **Figure 1** compares /ba/ and /fa/ between the original recordings and the vocoded versions.

Nonsense pictures. Nonsense pictures in the PA task were from the "fribble" image set (Databases/TarrLab/(http://wiki.cnbc.cmu.edu/Novel_Objects)). Fribbles comprise 12 species with distinct body "core" shape and color, with 81 exemplars per specie obtained by varying the forms of each of four appendage parts. From the available images, 13 lists of 12 images each were created such that each list used three different body forms and no duplicated appendage forms, rendering the images within each list highly distinctive (Williams and Simons, 2000). No appendage was repeated across lists.

Design

Figure 2 outlines the overall design of the experiment. Participants completed pre-training consonant identification familiarization and pre-training forced-choice consonant identification. Then, on each of four different days, they completed three blocks of PA training and AO testing associated with one word list. Participants were assigned to either AV or AO training for the duration of the experiment. Following the PA training and testing, participants were tested again on AO forced-choice consonant identification.

Consonant identification familiarization procedure. The pre- and post-training forced-choice consonant identification involved all the English consonants. Because English orthography is not uniquely mapped to English phonemes, participants were first familiarized with the orthographic transcription system, which

Table 3 | Word lists for paired-associates task. Lists 1–4 were used in Experiment 1.

Training list 1	Test list 1	Training list 2	Test list 2	Training list 3	Test list 3	Training list 4	Test list 4
sICUd	sICUd	mITak	mITak	hIluz	hIluz	kizxl	Kizxl
pcriD	pcriD	lRman	lRman	Cudxk	Cudxk	wEsIk	wEsIk
CRfIG	CRfIG	Sczxn	Sczxn	bUran	bUran	Bincl	Bincl
wInct	wInct	Bodut	Bodut	Jobxt	Jobxt	Pcgxs	Pcgxs
kUmxl	kUmxl	Ridap	Ridap	m@fis	m@fis	TuSxz	TuSxz
hUbIG	hUbIG	zEriC	zEriC	kcraC	kcraC	s@bad	s@bad
digaz	SEsxl	pIDRz	pEt∧f	tEfRk	zEnop	Yupan	m@d∧v
lIZxs	bozEn	wRsIG	f@Jxs	Ncrim	dik∧p	hob∧k	SRfxn
mcTxs	JovRs	k@fRt	viw∧s	ril∧n	yUS∧k	dISxp	l@kat
tETan	m@tuT	TEmat	nIsxJ	TIfxs	rIZxl	vIpxd	zESxm
rip∧J	fctab	dib∧J	JUkiz	fICUt	Lctak	m@Jxv	CIlxz
Yulat	D@zxk	sEJud	wEsxJ	S@dxz	w@vxt	Nupis	fEkRz

Training list 5	Test list 5	Training list 6	Test list 6	Practice list 1	Practice list 2
zudxn	Zudxn	mEzud	mEzud	fISxb	hRsak
wizcg	Wizcg	bikud	bikud	ballot	pEJun
m@nad	m@nad	SIzxv	SIzxv	yUtin	bUris
C@zxd	C@zxd	hivan	hivan	mRsaC	JEroz
pincg	Pincg	vid∧n	vid∧n	DEkxs	pEvxk
y@pat	y@pat	JIfxl	JIfxl	bon∧f	Mizcl
b@GIt	k@tup	nimat	pEriT	zErIp	dISus
hozIk	gIsan	pasIk	naSis	ripEs	dipcs
lipRt	h@Jus	rigab	kRCxm	hISxd	vRpad
fcris	Sigak	tcrab	gEsak	hon∧t	mivRd
nopiz	Fonab	k@pIG	wimun	hImut	dIs∧f
rik∧f	rEmRz	wilus	zIJxl	p@fxJ	wEvRz

*Practice List 1 was used to familiarize participants with the task. Lists 1–3 were used for AO training and testing, and Lists 4–6 for AV training and AO testing. Practice List 2 was presented AO, and Practice List 1 was presented AV. Test lists always show that the first six words in the list were carried into testing and six new words were substituted for six trained words. (**Table 2** gives the transcription key for phoneme mappings.)*

was compatible with single-character keyboard entry. An answer key (the consonants listed in **Table 2**), also available during testing, was used to explain the orthographic system. During familiarization, participants filled out two self-scored worksheets, one with the key available and one without. The participants' task was to transcribe 48 consonants in real English words while looking at the key and then 71 consonants in real words without looking at the key. A six-item practice test was randomly selected from two practice lists. All the participants were able to use the orthographic transcription system.

Pre- and post-training test procedure. Audio-only forced-choice consonant identification was carried out with CVCVC nonsense words. On each trial, following presentation of a stimulus, a response string of the form "__-__-__" appeared on the monitor, and the participants typed, in order, the three consonants that they had perceived in the AO spoken stimulus. They were instructed to guess when necessary. Only characters from the response set were displayed in the response string. It was possible to correct a response, and use of the enter key completed the trial. No feedback was given for the correctness of the responses. Different test lists were assigned across pre- and post-training testing, and list order was counter-balanced across participants.

Paired-associates training procedure. **Figure 3** outlines the design of a PA training trial. During training, the participant's task was to learn, with feedback over repeated presentations, lists of individual associations between 12 fribble images and 12 CVCVC vocoded spoken nonsense words. In **Figure 3**, an AV training trial is shown in the left column and an AO training trial is shown in the right column. Each trial began with a computer-monitor display of the 12-fribble image matrix (three rows of four columns, with image position within the matrix randomly selected on a trial-by-trial basis). During AV training, a video of the talker was played in synchrony with the spoken audio, and during AO training, a single still image of the talker's face was displayed on the monitor during audio presentation. The talker was presented on a different monitor than the fribble matrix monitor, and a large arrow appeared on the bottom of the fribble monitor pointing left to remind the participant to focus attention on the talker. The participant used the computer mouse to choose a fribble image following the speech stimulus. Feedback was given by outlining the correct fribble in green and an incorrect choice in red. After a short interval, the speech stimulus was always repeated, while the fribble images and borders remained unchanged. A training block comprised two repetitions of the 12 paired associations in pseudorandom order. Prior to the first training list in each

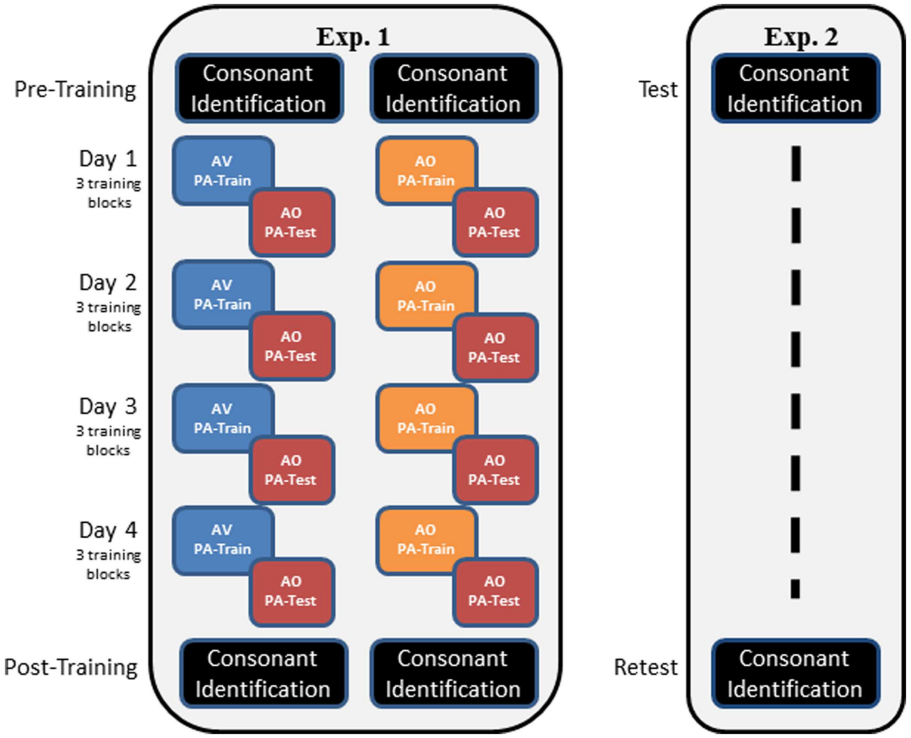


FIGURE 2 | Overall designs of Experiments 1 and 2. In Experiment 1, participants carried out pre-training consonant identification, followed by either AV or AO training on four stimulus lists, with AO tests on each list. Training of three blocks per list was carried out on a separate day for each list. Post-training, participants were tested again on AO forced-choice consonant identification. In Experiment 2, participants were tested only on forced-choice consonant identification on two different days (test, re-test). The two administrations of the forced-choice consonant identification used different stimulus lists.

condition (AV or AO), participants were given practice with one block of six trials.

Paired-associate testing procedure. paired-associates testing immediately followed training. The testing procedure was the same as that of PA training, except the stimuli were always AO, no feedback was given, the stimulus was not repeated during the trial, and each response triggered the next trial. Six of the trained spoken words and all 12 of the fribble images were used for testing. The associations for the six retained words were unchanged. Six new nonsense words were paired with the fribble images of the discarded words. A testing block comprised, in pseudorandom order, one presentation of the 12 stimuli, and three blocks were presented. The test score was the proportion of correct paired associations of trained words.

Apparatus. Audiovisual CVCVC tokens were digitized, edited, and conveyed to digital video disk (DVD) format. The acoustic waveforms were vocoded in real time, and the audio stimuli were output at a calibrated 65 dB A-weighted sound pressure level (SPL) using a JBL LSR6325P-1 loudspeaker. Participants were tested in an Industrial Acoustics Company (IAC) double-walled sound-attenuating booth using a standard computer interface that included a 51 cm LCD monitor, and a 35.6 cm Sony PVM-14N5U NTSC video monitor for display of speech video from the DVD. Monitors were located about 1 m from the participant's eyes, so that the computer-monitor subtended a visual angle of 23.1° horizontally and 17.3 vertically with the 12 fribble matrix filling the monitor. The visual speech was displayed on the NTSC monitor with the talker's head subtending visual angles of 3.9° horizontally and 5.7 vertically. Custom software was used to run the experiment.

Analyses. In order to stabilize the variance of proportion correct scores, the arcsin transformation, $X^1 = \sin^{-1}\sqrt{X}$ was computed, where X was the proportion correct score computed over the appropriate set of trials. All analyses were also conducted in parallel on untransformed scores, and all of the parallel analyses agreed. Statistics are reported on the arcsin transformed data, but tables, means, and figures are untransformed to facilitate interpretation.

Results and discussion

Paired-associates training. Initial inspection of the training and testing data showed there to be wide individual variation. There were participants who were unable to learn associations to an acceptably high-level of accuracy within the three training blocks. In order to assure that a relatively similar level of PA learning had taken place across training conditions, the criterion of at least 75% correct on the third training block was set for use of a participant's data. That is, we chose to remove the data sets obtained

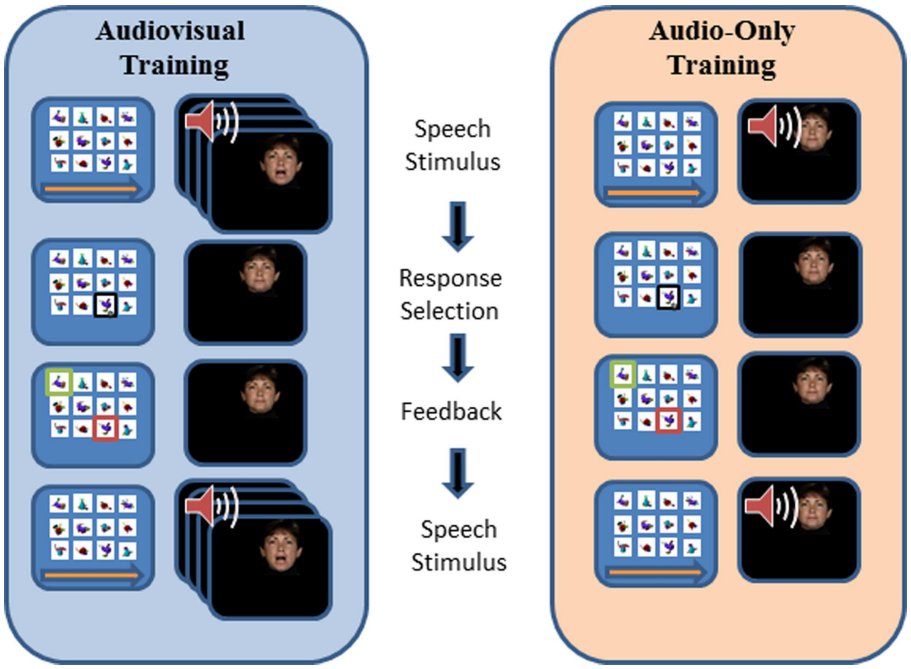


FIGURE 3 | Trial structure for paired-associates training. A speech stimulus was presented, followed by the participant's response selection, followed by feedback and a repetition of the speech stimulus. Each panel depicts the screen showing the fribble images side-by-side with the video monitor showing the talker. The trial structure for AV and AO training followed the same sequence, except that during AV training the video was played synchronously with the audio, and during AO training a still neutral face was played during the audio.

from participants who appeared to have difficulty learning associations *per se*. This criterion removed data from 10 participants from analyses. An additional participant was dropped because of scoring 6% correct on the test of one list, deviating greatly from typical test performance (mean = 94%, minimum = 67%, maximum = 100%). The analyses reported henceforth are on the data from 25 participants, 12 in the AV-trained group and 13 in the AO-trained group.

To examine performance during training, scores were submitted to RMANOVA with the within subjects factors of training list (1–4) and training block (1–3), and the between-subjects factor of training group (AO-trained, AV-trained). Importantly, no evidence was obtained for a reliable main effect or interaction with training group. Reliable main effects were obtained for training list $F(3, 69) = 19.26$, MSE = 0.49, $p < 0.001$, $\eta_p^2 = 0.46$, and training block, $F(2, 46) = 651.09$, MSE = 14.41, $p < 0.001$, $\eta_p^2 = 0.97$. A significant interaction between list and block (see **Table 4**), $F(6, 138) = 6.77$, MSE = 0.08, $p < 0.001$, $\eta_p^2 = 0.23$, was also obtained. **Table 4** shows that, with experience, learning was faster.

Table 4 | Experiment 1 training scores as a function of list and block.

	Block 1	Block 2	Block 3
List 1	31(2.0)	76(3.3)	95(1.3)
List 2	42(2.2)	90(2.0)	98(0.8)
List 3	49(2.5)	93(1.6)	96(1.2)
List 4	51(2.1)	91(1.8)	97(1.0)

The means are presented with the standard error of the mean in parenthesis.

Paired-associates test results

The critical question was whether the AV-trained participants were more accurate than AO-trained participants when both were tested with AO stimuli. The proportion correct PA test scores based on three repetitions of each of the six trained items was computed. The values were submitted to RMANOVA with the within subject factor of training list (1–4) and the between subject factor training condition (AO, AV). A main effect of training condition, $F(1, 23) = 7.619$, MSE = 0.36, $p < 0.05$, $\eta_p^2 = 0.25$, was obtained. The AV-trained participants had higher AO test scores (97% correct test scores, SE = 1.4) than did the AO-trained participants (92% correct test scores, SE = 1.4). No other effects were reliable. The responses to the six untrained words that were presented during testing were also checked for accuracy, and the scores were very low.

Pre- and post-training results

Forced-choice consonant identification data were collected pre- and post-training on independent lists of AO nonsense words. Proportion correct identification scores for consonants in initial, medial, and final position were computed separately on pre- and post-training data. Scores were submitted to RMANOVA with within-subject factors of time of testing (pre- versus post-training), consonant position (initial, medial, and final), and between-subjects factor group (AV-trained, AO-trained). The main effects of time of testing, $F(1, 23) = 141.08$, MSE = 0.98,

$p < 0.001$, $\eta_p^2 = 0.86$, and of consonant position, $F(2, 46) = 49.22$, MSE = 0.28, $p < 0.001$, $\eta_p^2 = 0.68$, were both reliable.

The interaction between time of testing and group was reliable, $F(1, 23) = 8.54$, MSE = 0.06, $p < 0.05$, $\eta_p^2 = 0.27$. The AV-trained participants had lower pre-training forced-choice consonant identification scores and higher post-training scores (AV-trained pre 32% correct, post 50% correct; AO-trained pre 35% correct, post 47% correct), improving on average by 18% points. The AO-trained participants group improved their scores on average by 12% points. Because the two groups were different at pre-training, as well as post-training, post-training − pre-training gain scores were computed and submitted to an independent samples *t-test*. The gains obtained by the AV-trained group were significantly larger than the gains of the AO-trained group, $t(23) = 2.91$, $p < 0.05$ (see **Figure 4**).

The interaction between time of testing and consonant position was reliable, $F(2, 46) = 4.49$, MSE = 0.02, $p < 0.05$, $\eta_p^2 = 0.16$ (see **Table 5**). *Post hoc* tests with RMANOVA using the results for the individual consonant positions (initial, medial, and final) revealed that the magnitude of the difference in accuracy between initial and medial consonants was larger post-training than pre-training, $F(1, 24) = 7.45$, MSE = 0.07, $p < 0.05$, $\eta_p^2 = 0.24$, as was the difference between final and medial consonants, $F(1, 24) = 5.67$, MSE = 0.07, $p < 0.05$, $\eta_p^2 = 0.19$. That is, the biggest perceptual learning gains were obtained for medial consonants (see **Figure 4**). AV-trained participants gained 24% points accuracy for medial consonants, and AO-trained participants gained 17% points.

EXPERIMENT 2 NO-TRAINING CONTROL

In Experiment 1, AV training resulted in better AO paired association learning and more accurate forced-choice consonant identification than did AO training. However, the design could not be used to conclude that all gains on the forced-choice consonant identification task were due to training. Therefore, a control experiment was conducted in which the forced-choice consonant identification task was administered twice but *without* intervening training.

Materials and methods

Subjects. Ten volunteers, aged 22–48 years of age, two male, participated in the experiment. The criteria for inclusion were the same as in Experiment 1.

Procedure. Only the brief AO consonant familiarization procedure, practice, pre-training (test), and post-training (re-test) consonant identification tests were administered (**Figure 2**). The time between test and re-test ranged from 3 to 16 days (mean = 8.1 days). The procedures for administering the forced-choice consonant identification were the same as in Experiment 1.

Results and discussion. The test and re-test forced-choice consonant identification data were submitted to RMANOVA with within-subject factors of time of testing (test, re-test) and consonant position (initial, medial, final). The main effects of time of testing, $F(1, 9) = 24.49$, MSE = 0.10, $p < 0.05$, $\eta_p^2 = 0.73$, and of consonant position, $F(2, 18) = 32.55$, MSE = 0.13, $p < 0.001$, $\eta_p^2 = 0.78$, were reliable. There were no reliable interactions. Identification accuracy increased from test (36% correct, SE = 2.7) to re-test (44% correct, SE = 3.1). Linear contrasts revealed that accuracy differed among all three consonant positions (initial = 34%, SE = 2.7; medial = 49%, SE = 3.6; final = 37% correct, SE = 2.7) (see **Table 5**).

Consonant identification gain scores from Experiments 1 and 2 (**Figure 4**) were submitted to RMANOVA with the between subject factor training group (AO-trained and AV-trained from Experiment 1 and no-training control from Experiment 2) and the within subject factor consonant position (initial, medial, final).

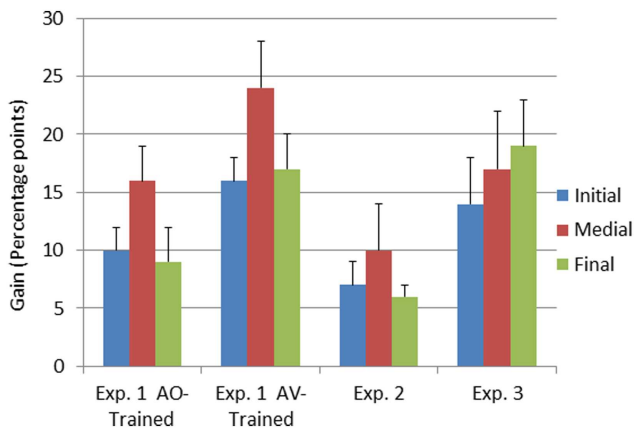


FIGURE 4 | Pre-to-post-training gain scores as a function of experiment and consonant position. Gain scores represent the means of the arithmetic difference between first and second forced-choice consonant identification test scores obtained in Experiments 1–3. The error bars represent 1 SE of the mean. Results are shown separately for the three consonant positions in the CVCVC stimuli.

Table 5 | Pre-training and post-training forced-choice consonant identification scores across experiments as a function of consonant position.

		Consonant Position		
		Initial	**Medial**	**Final**
Experiment 1				
AO training	Pre-	30 (1.7)	41 (3.7)	34 (2.5)
	Post-	40 (2.2)	58 (3.2)	43 (3.0)
AV training	Pre-	27 (1.7)	37 (3.9)	30 (2.6)
	Post-	43 (2.3)	61 (3.3)	47 (3.1)
Experiment 2	Test	31 (3.2)	44 (3.4)	34 (2.5)
	Re-test	37 (2.5)	54 (4.7)	40 (3.1)
Experiment 3	Pre-	31 (2.4)	47 (4.4)	34 (2.6)
	Post-	46 (4.4)	64 (4.5)	53 (4.1)

The tabled values are the percent correct means and standard error of the means in parentheses for each of the consonant positions in the CVCVC stimuli. In Experiments 1 and 3, the scores were obtained pre- and post-training. In Experiment 2, the scores were obtained without intervening training (test, re-test).

Training group was a reliable factor, $F(2, 32) = 10.42$, MSE = 0.13, $p < 0.001$, $\eta_p^2 = 0.83$. Pair-wise comparisons between AO-trained (Experiment 1), AV-trained (Experiment1), and the no-training control (Experiment 2) showed that AV-trained participants had significantly higher forced-choice consonant identification gain scores than controls (see **Figure 4**) ($p < 0.05$). But gain scores of Experiment 1 AO-trained participants were not reliably different from those of the no-training controls. Thus, across experiments, only the AV-trained participants demonstrated auditory perceptual learning that was more successful than merely participating in a test-re-test consonant forced-choice identification task.

Consonant position was reliable in the comparison across groups, $F(2, 64) = 4.37$, MSE = 0.04, $p < 0.05$, $\eta_p^2 = 0.12$. Pair-wise comparisons revealed that medial pre-to-post gain scores differed from initial and final gain scores (initial = 11.6%, SE = 1.3; medial = 17.6%, SE = 2.3; final = 11.2%, SE = 2.7; $p < 0.05$).

EXPERIMENT 3 WITHIN-PARTICIPANT AUDIOVISUAL AND AUDITORY-ONLY TRAINING

In Experiment 3, a modified training protocol was carried out in order to test whether the AV training advantage in Experiment 1 would be reliable under a different training protocol. Training followed that of Experiment 1, except that participants were trained until they reached the criterion of 92% correct within a training block and list. Also, AV and AO training conditions were alternated across lists, and six lists were trained (**Figure 5**).

Materials and methods

Subjects. Fifteen participants were recruited and started the experiment. The criteria for inclusion in the experiment were the same as in Experiment 1. Two dropped out due to difficulty learning the paired associations. The 13 who completed testing were ages 21–51 years (mean = 28 years), with two males.

Procedures. Mixed PA AV and AO training was given with counter-balanced initial condition and six lists total (AO, AV, AO, AV, AO, AV, or AV, AO, AV, AO, AV, AO) (see **Figure 5**). Testing was always AO. Every list of paired associations was trained until the participant scored at least 92% correct. Then, in the same session, the corresponding AO test was administered. Participants were permitted to train on more than one list per session. The forced-choice consonant identification test was administered pre- and post-training as in Experiment 1.

Results

Paired-associates training. The number of training trials to achieve the 92% correct criterion was submitted to RMANOVA with the within subjects factors of training condition (AO, AV) and list (first, second, third). The main effect of list, $F(2, 24) = 4.85$, MSE = 1602.46, $p < 0.05$, $\eta_p^2 = 0.29$, was the only factor that reached significance. Pair-wise comparisons indicated that, across training condition, more trials (mean = 76.6, SE 6.16) were needed to reach criterion on the first list than on the second (mean = 64.6, SE 5.18) and third (mean = 61.8, SE 5.74) ($p < 0.05$), and the latter two did not differ.

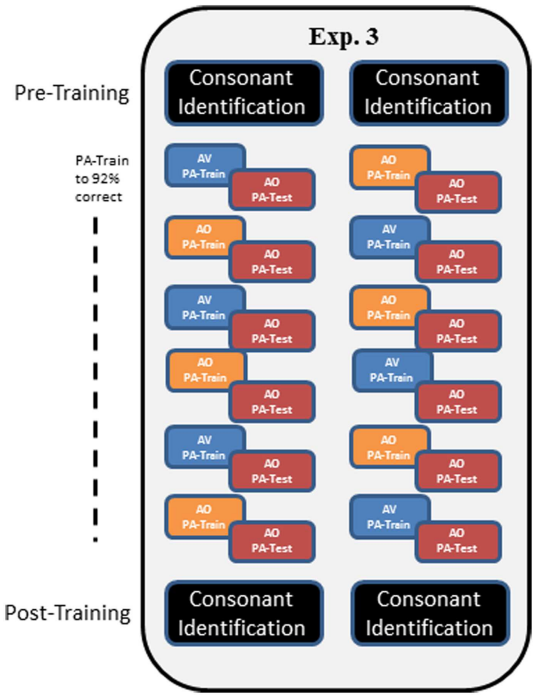


FIGURE 5 | Overall design of Experiment 3. In Experiment 3, participants carried out pre-training consonant identification, followed by alternating AV and AO training, counter-balanced across participants as shown by the two columns of train versus test in the figure. Training blocks on each list were repeated until the participant achieved 92% correct. Then AO testing was administered. Post-training, participants were tested again on AO forced-choice consonant identification.

The mean accuracy scores over the blocks to criterion within a list were also submitted to RMANOVA with the within subjects factors of training condition (AO, AV) and list (first, second, third). Again, the main effect of list, $F(2, 24) = 14.15$, MSE = 0.04, $p < 0.001$, $\eta_p^2 = 0.54$, was the only significant factor. Pair-wise comparisons indicated that the first list was less accurate (mean = 66.5, SE = 1.5) than the second (mean = 71.6, SE = 1.8), which was less accurate than the third (mean = 73.9, SE = 1.2; $p < 0.05$).

Paired-associates test results. The PA test results were submitted to RMANOVA with within subject factors of training condition (AO, AV) and list (first, second, third). The main effect of training condition was the only significant effect, $F(1, 12) = 8.44$, MSE = 0.25, $p < 0.05$, $\eta_p^2 = 0.41$. AO-trained PA test scores were higher (94.0% correct mean test score, SE = 1.8) than AV-trained PA test scores (88.9% correct mean test score, SE = 2.5).

In Experiment 1, AV PA training resulted in higher AO test scores (97% correct test scores, SE = 1.4) than did AO training (92% correct AO test scores, SE = 1.4). To compare PA test scores across Experiments 1 and 3 (which had different designs), we pooled test scores within subject separately for AV- and AO-trained lists in each experiment. The results showed that AV training in Experiment 1 was significantly more effective than in Experiment

3, t (23) = 2.78, $p < 0.05$. But the AO scores were not different across experiments.

The discrepancy in PA results across Experiments 1 and 3 might have been related to the different criteria for learning that was used to accept data. In Experiment 1, a performance criterion of 75% correct on the third training block for each list was used for inclusion of data. This resulted in dropping 10 out of 36 participants (another one was dropped for an exceptionally low AO test score on trained stimuli). In Experiment 3, two participants were unable to learn the PA stimuli to criterion of 92% correct. However, if we had imposed the 75% correct criterion on the third training block in Experiment 3, 4 out of 13 participants would have failed, which is a comparable proportion to that of Experiment 1. Thus, the results across experiments seem unlikely to be related to group differences in ability to learn paired associations.

Pre- and post-training consonant identification. Forced-choice consonant identification scores were submitted to RMANOVA with the within subjects factors of time of testing (pre- versus post-training) and consonant position (initial, medial, final). The main effects of time of testing, $F(1, 12) = 15.83$, MSE = 0.68, $p < 0.05$, $\eta_p^2 = 0.57$, and of consonant position, $F(2, 24) = 38.99$, MSE = 0.23, $p < 0.001$, $\eta_p^2 = 0.77$, were reliable (see **Figure 4** and **Table 5**). The interaction between time of testing and consonant position was not reliable. Consonant identification accuracy increased from pre- (37% correct, SE = 2.7) to post-training (54% correct, SE = 4.1). Linear contrasts revealed that accuracy differed between all three positions (initial = 38%, SE = 2.7; medial = 56%, SE = 3.8; final = 43% correct, SE = 2.7).

GENERAL DISCUSSION

The results of this study suggest that AV training can promote auditory perceptual learning of novel, vocoded speech more effectively than AO training. But the training procedure affects perceptual learning outcomes. In Experiment 1, PA training was carried out with disyllabic spoken nonsense words and nonsense pictures. Participants were assigned to learn the associations with either AV or AO speech stimuli within a fixed number of trials. AV training was significantly more effective than AO training, as measured by testing how well the paired associations could be identified with AO stimuli. Pre- and post-training forced-choice consonant identification was also administered AO with untrained sets of disyllabic spoken nonsense words. On this task also, AV-trained participants were more accurate than AO-trained participants. Perception of medial consonants was significantly affected by AV training. AV-trained participants gained 24% points accuracy for medial consonants, and AO-trained participants gained 17% points. In Experiment 2, a control experiment, participants were tested twice in the forced-choice consonant identification paradigm but without intervening training or feedback of any kind. Their re-test scores were significantly higher than their initial scores. The consonant identification scores were then compared across Experiments 1 and 2. The comparison showed that AO-trained participants in Experiment 1 were *no more* accurate on consonant identification than re-tested participants in Experiment 2. In contrast, AV-trained participants in Experiment 1 were *more* accurate than re-test participants in Experiment 2. Experiment 3 was carried out using PA training that alternated between AV and AO conditions on a list-by-list basis (mixed training). Training was to a 92% correct criterion, and two more lists were trained than in Experiment 1. Lists tested after AO training resulted in significantly higher AO PA scores than lists tested after AV training. Test scores on the paired associations were compared across Experiments 1 and 3. AV-trained participants in Experiment 1 were significantly more accurate (97% correct) than participants in Experiment 3 following AV training (88.9% correct). AO-trained participants in Experiment 1 performed similarly to participants in Experiment 3 following AO training (Experiment 1, 92% and Experiment 3, 94.0% correct).

REVERSE HIERARCHY THEORY FOR MULTISENSORY SPEECH PROCESSING

The results of Experiment 1 suggest that multisensory stimuli can be used for improving unisensory perceptual learning. But the results of Experiment 3 suggest that multisensory stimuli can also impede unisensory perceptual learning. A theory of perceptual learning (Goldstone, 1998) is needed to explain these discrepant results. We have adopted the reverse hierarchy theory (RHT) of perceptual learning (Ahissar and Hochstein, 1997; Ahissar et al., 2008), because it attempts to explain perception and perceptual learning within the context of neural processing.

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

According to RHT, immediate perception relies on established high-level representations in the bottom-up sensory-perceptual pathway. When a new perceptual task needs to be carried out, naïve performance is initiated on the basis of immediate high-level perception. However, if the task cannot be readily performed with the existing mapping of low-level to high-level representations, and/or if there is incentive to increase the efficiency of task performance, then perceptual learning is needed. According to RHT, perceptual learning is the access to and remapping of lower-level input representations to higher-level representations. To carry out the remapping, perceptual learning involves "perception with scrutiny." That is, a backward search must be initiated to access the representational level of the information needed to carry out the perceptual task. A new mapping can then be made. Mapping changes can occur in both convergence and divergence patterns (Jiang et al., 2007b; Kral and Eggermont, 2007; Ahissar et al., 2008). That is, dissimilar lower-level input representations can map to the same higher-level representations; and similar lower-level input representations can map to different higher-level representations.

SPEECH PROCESSING PATHWAYS

Reverse hierarchy theory has not, to our knowledge, previously been extended to an explicit theory of multisensory constraints on

unisensory perceptual learning, but the evidence on the diversity and extent of cortical and subcortical multisensory connections (Foxe and Schroeder, 2005; Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008; Kayser et al., 2012) suggests that higher-level representations in one sensory-perceptual system can be used to gain access to lower-level representations in another sensory-perceptual system. **Figure 6** is a schematic view of auditory and visual speech processing pathways. It suggests that at each level of stimulus processing – basic features (e.g., spectrotemporal auditory features and spatiotemporal visual features not specific to speech), phonetic features (linguistically relevant sub-phonemic integrated basic features), phonemes (syllables or word forms, i.e., linguistically relevant categories) – there is the possibility of multisensory integrative processes and also unisensory representations. Various experimental results have been interpreted as evidence that visual speech information can converge as early as primary auditory cortex (e.g., Sams et al., 1991; Calvert et al., 1997; Giard and Peronnet, 1999; Möttönen et al., 2002; Raij et al., 2010), and anatomical animal studies have provided evidence of multisensory connectivity as low as primary visual and auditory areas (Ghazanfar et al., 2008; Falchier et al., 2012). Such results have been interpreted as support for early and obligatory multisensory integration (Rosenblum, 2008). Other findings point to multisensory integration at higher cortical levels, such as superior temporal sulcus, suggesting that extensive unisensory integration has occurred prior to integrative activity (Miller and D'Esposito, 2005; Hasson et al., 2007; Bernstein et al., 2008a; Nath and Beauchamp, 2011).

Figure 6 shows a parallel structure for unisensory auditory and visual speech processing. The parallel unisensory hierarchy for visual speech receives diverse support in the literature. For example, dissimilarity measures of visual speech stimuli significantly account for consonant perceptual dissimilarity (Jiang et al., 2007a; Files and Bernstein, in preparation). That is, physical optical measures can account for significant variance in visual perceptual identification and discrimination. Patterns of confusions for lipreading words are reliably accounted for by visual perception of spoken phonemes (Mattys et al., 2002). Visual perceptual confusions account for results on visual spoken word identifications better than auditory perceptual confusions (Auer, 2002). Visual speech mismatch negativity event-related potentials have been localized posterior to auditory temporal cortices (Ponton et al., 2009; Files and Bernstein, submitted), and visual speech processing has been localized with functional magnetic resonance imaging in posterior superior temporal cortex and adjacent middle temporal cortex, consistent with speech representation in the high-level vision pathway (Bernstein et al., 2011).

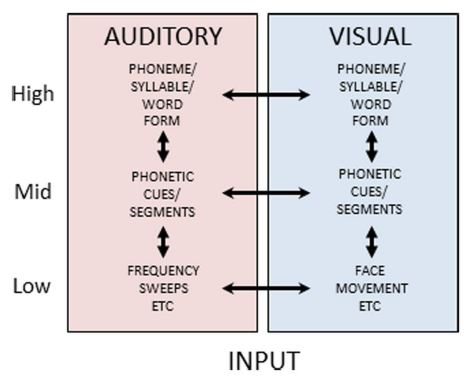


FIGURE 6 | Auditory and visual speech pathways. Figure 6 schematically depicts cortical processing pathways and their interactions for auditory and visual speech. Separate uni-sensory processing pathways are thought to exist for mapping from low-level stimulus features to high-level form-based representations of speech stimuli. During perceptual processing, information is thought to predominantly flow from low-to-high, however feedback pathways are available along both pathways. Additionally, at each level double arrowed lines between the pathways indicate the potential for multisensory integrative processing.

Thus, speech perception can be multisensory, visual-only, or auditory-only, and there is support for representations that correspond to these three possibilities. It also seems reasonable to conclude across the many results on speech perception involving auditory and visual stimuli that multisensory integration is available at every level of speech processing, consistent with a highly multisensory cerebral cortex (Ghazanfar and Schroeder, 2006). How could this diversity of integrative resources contribute to the discrepant results of Experiments 1 and 3?

EXPLANATION FOR DIVERGENT MULTISENSORY TRAINING OUTCOMES

In order to explain our divergent results, we need to focus on the level at which auditory perceptual learning took place. Our results point to phonetic features, which are linguistically relevant sub-phonemic representations that typically are said to map to phoneme categories (for discussion of features, Jakobson et al., 1961; Chomsky and Halle, 1968) but could also map directly to syllable, morpheme, or word-level categories (Grossberg et al., 1997; Vitevitch and Luce, 1999; Norris et al., 2000). The results point to auditory perceptual learning of phonetic features, because learning generalizes to forced-choice consonant identification in new words, and learning is differentially affected by the position of the consonant. If consonants were learned as unanalyzed units, we would not expect that their position in the word would be a significant effect in our results. The medial consonant affords the most phonetic feature information, which is obtained from the vowel transitions into and out of the consonant (Stevens, 1998), and therefore phonetic feature learning should result in more gains when feature information is richer. In addition, the largest amount of auditory learning was for the medial consonant position following AV training: Auditory perceptual learning was more sensitive to phonetic details in the auditory stimuli when the training was AV.

To be clear, phonetic features are integrated representations based on basic sound features. That phonetic features are complex combinations of information about the acoustic attributes of speech has been extensively researched (Stevens, 1998). For example, the place of articulation (e.g., involved in the distinction /b/ versus /d/) is instantiated in the acoustic signal partly by the center frequency and transitions of the speech formants (resonance of the vocal tract). The feature known as voicing (e.g., involved in the distinction /b/ versus /p/) is instantiated partly by

the temporal offset difference between consonant initiation in the supralaryngeal vocal tract and the onset of glottal pulsing (Lisker et al., 1977). Relatively little research has been carried out on the neural bases of phonetic feature processing, with most speech perception research focused on levels either lower than or higher than phonetic features (Binder et al., 2000; Scott, 2005; Hickok and Poeppel, 2007; Liebenthal et al., 2010), however, Obleser and Eisner (2009) have identified a site of phonetic feature processing anterior to the primary auditory cortical areas in superior temporal gyrus. This gives support to the possibility of focused phonetic feature learning.

When speech is degraded or transformed, perceptual confusions among phonemes can be described in terms of loss of phonetic feature distinctions (Miller and Nicely, 1955; Wang and Bilger, 1973). The problem for auditory perceptual learning of vocoded speech is to remap available basic auditory features (such as frequency and temporal features) in the novel transformation to phonetic features that support the perception of syllables, morphemes, and/or words.

Figure 7 illustrates our proposed model for the outcomes of Experiments 1 and 3 within the context of multisensory and unisensory processing resources and the RHT of perceptual learning. In **Figure 7**, the blue and red circles represent visual and auditory phonetic speech features, respectively. For purposes here and in **Figure 7**, the category that phonetic features target is not important to define, because the results of the three experiments point to auditory perceptual learning at the phonetic feature level targeting phonemes, and as pointed out above features could target phonemes, syllables, morphemes, or words.

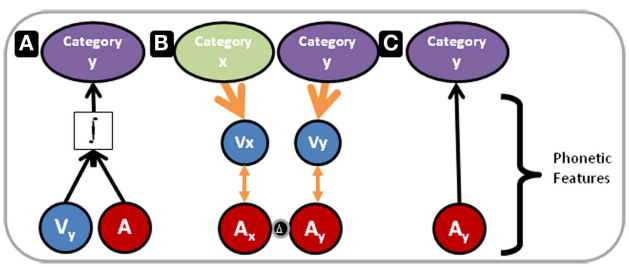


FIGURE 7 | Perceptual learning versus integration model. The blue and red circles in the lower part of Figure 7 represent visual and auditory phonetic speech features, respectively. These correspond to the mid level of processing in **Figure 6**. The categories at the top of the figure correspond to representations at the high-level of processing in **Figure 6**. **(A)** Depicts processing under conditions in which acoustic phonetic features alone are not sufficient to specify the phoneme category. The integrated audiovisual phonetic features do provide adequate information. Perceptual processing flows bottom-up, and remapping along the auditory pathway has not occurred. In contrast, **(B)** Depicts a reverse flow of information. As in **(A)**, Combined audiovisual information is sufficient to specify phoneme categories (not shown). However, here a reverse search is initiated. Higher-level visual speech categories, *x* and *y*, feed back to visual phonetic features, *Vx* and *Vy*, that use natural audiovisual correlations (orange double arrowed lines) to guide the search for relevant distinctions in acoustic-phonetic feature representations. The two red circles separated by a delta are labeled *Ax* and *Ay* because the acoustic phonetic features are now distinct. **(C)** Depicts auditory-only processing, following the perceptual learning depicted in **(B)**. The acoustic phonetic features alone are now sufficient to specify the phoneme category.

In **Figure 7A**, vocoding has removed or distorted the basic auditory information that is typically mapped to phonetic features of natural speech. The phonetic feature level is inadequate to specify the phoneme category (phoneme categories for purposes here). But the visual speech information provides the needed phonetic information (Summerfield, 1987), the information is integrated, and the perceptual task is carried out at an immediate high-level of perception, as predicted by RHT. However, with early integration the perceptual task can be accomplished without scrutiny of auditory lower-level representations, and if the visual stimulus is unavailable performance drops. This is our explanation for the finding in Experiment 3, in which performance following AV training was lower than following AO training.

Several factors in Experiment 3 could have reduced the likelihood that participants focused on the auditory information when the training was AV. RHT predicts that when semantic processing is required, low-level access is precluded (Ahissar et al., 2008; Nahum et al., 2008). In Experiment 3, participants were trained to criterion, and they were free to train on as many lists as possible during a training session. Trying to learn more than one list in a day could have directed attention to semantic relationships. Training to criterion on more than one list could have encouraged less attention to the auditory input, because it might have led participants to put a premium on the rate at which the paired associations were learned rather than on the accuracy of the AO tests. Also, given that perception of AV speech stimuli is frequently faster and more reliable (Sumby and Pollack, 1954; Van Wassenhove et al., 2005; Ross et al., 2007), we surmise that in Experiment 3 the perceived effort to learn the paired associations was lower under AV versus AO conditions. This perceived reduced effort might have also favored relying on high-level representations that were fed by AV integration. While it is true that semantic category training can result in retuning representations (Jiang et al., 2007b) and change in sensitivity to category boundaries (Goldstone, 1994), such training typically involves less diverse stimuli than the ones in the present study.

Figure 7B has two columns. Each has a downward arrow from a higher-level of visual speech category representation to a level that is correlated with auditory representations. Remapping from basic sound to phonetic features has taken place due to top-down guidance within the visual system. The red circles are labeled *Ax* and *Ay*, because phonetic features are now distinct. We think that the auditory distinctions that were learned in our study *must* be readily available at the level of basic features (not indicated in **Figure 7**), because learning was relatively fast and low-level auditory retuning is likely not affected over such a brief period (Kral and Eggermont, 2007). Likewise, the rapid learning argues against learning based on new connections via dendritic growth and arborization.

We hypothesize that this remapping process makes use of natural correlations between auditory and visual speech stimuli, indicated in **Figure 7B** with the double pointed arrows. These natural AV correlations provides a link whereby visual information can help guide attention to the relevant distinctions in the

auditory representations. Research on the predictability of acoustic signals from optical signals and *vice versa* has shown that there are high-levels of correlation between acoustic and optical speech signals (Yehia et al., 1998; Jiang et al., 2002; Jiang and Bernstein, 2011). Perceptual evidence shows that quantified correlation of the physical acoustic and optical speech signals can account for AV speech responses with matched and mismatched (McGurk type) stimuli (Jiang and Bernstein, 2011). Visual speech stimuli have been suggested to modify auditory speech processing through modulatory effects on neuronal excitability (Schroeder et al., 2008). Speech-in-noise experiments suggest that perceivers adjust their perception and neural networks change in relationship to the relative reliability of auditory or visual information (Ross et al., 2007; Nath and Beauchamp, 2011), or the temporal alignment of the stimuli (Miller and D'Esposito, 2005). We are suggesting that top-down processing from visual speech representations can guide access to distinctive auditory features that can be remapped to phonetic features for novel speech transformations. Top-down guidance via orthographic representations has been suggested as another basis for auditory perceptual learning of vocoded speech (Davis et al., 2005).These two types of top-down guidance might result in different learning. Specifically, the multisensory speech correlations might provide more fine-grained guidance for phonetic learning than orthography.

In **Figure 7C**, following the successful remapping, when AO stimuli are presented, the auditory mapping to the category is sufficient to carry out the task. **Figure 7C** corresponds to the result in Experiment 1 that AV PA training was more effective than AO training or merely re-testing in Experiment 2.

SOME IMPLICATIONS FOR TRAINING

Results reported here could be important clinically, for example, to crafting strategies for patients newly fitted with a cochlear implant (Zeng et al., 2004). The goal of such training is to assist the cochlear implant user in gaining access to the information in the degraded or impoverished signal delivered by the auditory prosthesis. Such patients can benefit from auditory training, but the benefits are typically not large (Fu et al., 2005; Stacey et al., 2010). A focus in training studies has been on which linguistic units such as phonological features, syllables, words, or sentences might best promote auditory perceptual learning (Fu et al., 2005; Stacey et al., 2010). However, the goals of training might be better served by focusing on the flow of information processing, specifically, the possibility that reverse hierarchy processing is needed to gain access to the available information (Kral and Eggermont, 2007; Auer and Bernstein, 2012). Focus is needed on the possibility that top-down guidance must be crafted that allows access to the level of representation where additional cues are available to be remapped. The current results support this view. But knowledge is also needed to predict when AV integration can impede auditory perceptual learning.

The results here are particularly relevant to training young cochlear implanted children who have not yet learned to read. In contrast to literate normal-hearing adults who can use orthographic representations or clear speech to guide perceptual learning (Davis et al., 2005; Hervais-Adelman et al., 2011), children's guides are often limited to multisensory information delivered via lipreading, visual signed language or fingerspelling, and/or vibrotactile speech displays (Bernstein et al., 1991; Auer et al., 1998).

A concerted effort was made in the twentieth century to design and test vibrotactile speech perception prostheses to supplement lipreading by deaf individuals including children. While the intent of the research was to learn how to convey speech through mechanical vibration signals, combined visual-vibrotactile training was shown to be associated with improved visual-only speech perception (Boothroyd and Hnath-Chisolm, 1988; Eberhardt et al., 1990; Bernstein et al., 1991; Kishon-Rabin et al., 1996). These improvements in lipreading sometimes exceeded the vibrotactile learning. This type of result suggests that when a novel speech signal is combined with a more familiar one, attention might be directed toward discerning additional information from the more familiar signal rather than the target novel signal. Indeed, in a companion study (in preparation) to this one on prelingually deaf adults who obtained cochlear implants as adults, we found that AV training resulted in faster PA learning but poorer auditory-only test scores, consistent with attention to and reliance on the more familiar visual stimuli. Indeed, there is evidence that visual perceptual abilities and multisensory integration are affected by cochlear implant usage in adults (Rouger et al., 2007). Understanding is needed for how to devise training that uses multisensory stimuli to guide unisensory perceptual learning, rather than only effecting immediate high-level perception with concomitant failure to achieve discernment of available low-level distinctions.

SUMMARY AND CONCLUSION

In summary, the results reported here do not fall under the rubrics of faster or more accurate AV versus AO speech perception, effects that have been well-documented (e.g., Sumby and Pollack, 1954; Bernstein et al., 2004; Van Wassenhove et al., 2005; Ross et al., 2007). They concern AV versus AO training effects on auditory-only perceptual learning. The information in a visual speech stimulus, presented in synchrony with a correlated but degraded auditory stimulus, can be effective in promoting auditory speech perceptual learning of the degraded stimuli. The visual information can promote more learning than the auditory stimuli alone, because of the correlations between auditory and visual features or cues, and because top-down visual processes can guide access to available but unused auditory cues. However, the multisensory speech stimuli typically are more informative and easier to perceive, and multisensory perception can rely on integrated representations, thereby possibly impeding unisensory perceptual learning. Research is needed on what perceptual learning procedures are required so that multisensory stimuli can be used reliably to enhance unisensory perceptual learning.

ACKNOWLEDGMENTS

We thank our test subjects for their participation and our technicians for supporting the data acquisition. Research supported by NIH/NIDCD DC008308.

REFERENCES

Ahissar, M., and Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature* 387, 401–406.

Ahissar, M., Nahum, M., Nelken, I., and Hochstein, S. (2008). Reverse hierarchies and sensory learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 285–299.

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

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

Auer, E. T. Jr., and Bernstein, L. E. (2012). "Plasticity for multisensory speech communication: evidence from deafness and normal hearing," in *The New Handbook of Multisensory Processing*, ed. B. E. Stein (Cambridge, MA: MIT), 453–466.

Auer, E.T. Jr., Bernstein, L. E., and Coulter, D. C. (1998). Temporal and spatio-temporal vibrotactile displays for voice fundamental frequency: an initial evaluation of a new vibrotactile speech perception aid with normal-hearing and hearing-impaired individuals. *J. Acoust. Soc. Am.* 104, 2477–2489.

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.

Bernstein, L. E. (2012). "Visual speech perception," in *AudioVisual Speech Processing*, eds E. Vatikiotis-Bateson, G. Bailly, and P. Perrier (Cambridge: Cambridge University), 21–39.

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

Bernstein, L. E., Auer, E. T. Jr., Wagner, M., and Ponton, C. W. (2008a). Spatio-temporal dynamics of audiovisual speech processing. *Neuroimage* 39, 423–435.

Bernstein, L. E., Lu, Z. L., and Jiang, J. (2008b). Quantified acoustic-optical speech signal incongruity identifies cortical sites of audiovisual speech processing. *Brain Res.* 1242, 172–184.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., and O'connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: performance of normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* 90, 2971–2984.

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

Bernstein, L. E., Jiang, J., Pantazis, D., Lu, Z.-L., and Joshi, A. (2011). Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Hum. Brain Mapp.* 32, 1660–1667.

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

Boothroyd, A., and Hnath-Chisolm, T. (1988). Spatial, tactile presentation of voice fundamental frequency as a supplement to lipreading: results of extended training with a single subject. *J. Rehabil. Res. Dev.* 25, 51–56.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 593–596.

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.

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

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical informatio drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *J. Exp. Psychol. Gen.* 134, 222–241.

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

Eberhardt, S. P., Bernstein, L. E., Demorest, M. E., and Goldstein, M. H. Jr. (1990). Speechreading sentences with single-channel vibrotactile presentation of voice fundamental frequency. *J. Acoust. Soc. Am.* 88, 1274–1285.

Falchier, A., Cappe, C., Barone, P., and Schroeder, C. E. (2012). "Sensory convergence in low-level cortices," in *The New Handbook of Multisensory Processing*, ed. B. E. Stein (Cambridge, MA: MIT), 67–79.

Falchier, A., Renaud, L., Barone, P., and Kennedy, H. (2001). Extensive projections from the primary auditory cortex and polysensory area STP to peripheral area V1 in the macaque. *Abstr. Soc. Neurosci.* 27.

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47.

Foxe, J. J., and Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport* 16, 419–423.

Fu, Q.-J., Galvin, J., Wang, X., and Nogaki, G. (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoust. Res. Lett. Online* 6, 106–111.

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.

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.

Ghazanfar, A. A., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci. (Regul. Ed.)* 10, 278–285.

Giard, M. H., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 11, 473–490.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *J. Exp. Psychol. Hum. Percept. Perform.* 123, 178–200.

Goldstone, R. L. (1998). Perceptual learning. *Annu. Rev. Psychol.* 49, 585–612.

Green, K. P., and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Percept. Psychophys.* 45, 34–42.

Grossberg, S., Boardman, I., and Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 483–503.

Hasson, U., Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2007). Abstract coding of audiovisual speech: beyond sensory representation. *Neuron* 56, 1116–1126.

Hazan, V., Sennema, A., Faulkner, A., and Ortega-Llebaria, M. (2006). The use of visual cues in the percpetion of non-native consonant contrasts. *J. Acoust. Soc. Am.* 119, 1740–1751.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., Taylor, K. J., and Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 293–295.

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

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Iverson, P., Bernstein, L. E., and Auer, E. T. Jr. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Commun.* 26, 45–63.

Jakobson, R., Fant, C. G. M., and Halle, M. (1961). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.* Cambridge, MA: MIT.

Jiang, J., Alwan, A., Keating, P., Auer, E. T. Jr., and Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP J. Appl. Signal Processing* 2002, 1174–1188.

Jiang, J., Auer, E. T. Jr., Alwan, A., Keating, P. A., and Bernstein, L. E. (2007a). Similarity structure in visual speech perception and optical phonetics. *Percept. Psychophys.* 69, 1070–1083.

Jiang, X., Bradley, E. D., Rini, R. A., Zeffiro, T., Vanmeter, J., and Riesenhuber, M. (2007b). Categorization training results in shape- and category-selective human neural plasticity. *Neuron* 53, 891–903.

Jiang, J., and Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psychol.*

Hum. Percept. Perform. 37, 1193–1209.

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799.

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2009). Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. *Hear. Res.* 258, 80–88.

Kayser, C., Petkov, C. I., Remedios, R., and Logothetis, N. K. (2012). "Multisensory influences on auditory processing: perspectives from fMRI and electrophysiology," in *The Neural Bases of Multisensory Processes*, eds M. M. Murray and M. T. Wallace (Boca Raton, FL: CRC), 99–113.

Kishon-Rabin, L., Boothroyd, A., and Hanin, L. (1996). Speechreading enhancement: a comparison of spatial-tactile display of voice fundamental frequency (F0) with auditory F0. *J. Acoust. Soc. Am.* 100, 593–602.

Kral, A., and Eggermont, J. J. (2007). What's to lose and what's to learn: development under auditory deprivation, cochlear implants and limits of cortical plasticity. *Brain Res. Rev.* 56, 259–269.

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

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cereb. Cortex* 20, 2958–2970.

Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., and Mandler, R. (1977). On pushing the voice onset-time (VOT) boundary about. *Lang. Speech* 20, 209–216.

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19 , 577–621.

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.

Mattys, S. L., Bernstein, L. E., and Auer, E. T. Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Percept. Psychophys.* 64, 667–679.

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

Miller, G. A., and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 301–315.

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.

Möttönen, R., Krause, C. M., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cogn. Brain Res.* 13, 417–425.

Nahum, M., Nelken, I., and Ahissar, M. (2008). Low-level information and high-level perception: the case of speech in noise. *PLoS Biol.* 6:e126. doi:10.1371/journal.pbio.0060126.

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.

Nath, A. R., and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787.

Norris, D., McQueen, J. M., and Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behav. Brain Sci.* 23, 299–370.

Obleser, J., and Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends Cogn. Sci. (Regul. Ed.)* 31, 14–19.

Ponton, C. W., Bernstein, L. E., and Auer, E. T. Jr. (2009). Mismatch negativity with visual-only and audiovisual speech. *Brain Topogr.* 21, 207–215.

Raij, T., Ahveninen, J., Lin, F. H., Witzel, T., Jaaskelainen, B. L., Israeli, E., et al. (2010). Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *Eur. J. Neurosci.* 31, 1772–1782.

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-reading*, eds B. Dodd and R. Campbell (London: Lawrence Erlbaum), 97–113.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Curr. Dir. Psychol. Sci.* 17, 405–409.

Ross, L. A., Saint-Amour, D., Leavitt, V. N., Javitt, D. C., and Foxe, J. J. (2007). Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., and Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7295–7300.

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

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145.

Schroeder, C. E., and Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cogn. Brain Res.* 14, 187–198.

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.

Scott, S. K. (2005). Auditory processing – speech, space and auditory objects. *Curr. Opin. Neurobiol.* 15, 197–201.

Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123(Pt 12), 2400–2406.

Scott, S. K., Rosen, S., Lang, H., and Wise, R. J. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech – a positron emission tomography study. *J. Acoust. Soc. Am.* 120, 1075–1083.

Seitz, P. F., Bernstein, L. E., Auer, E. T. Jr., and Maceachern, M. (1998). *PhLex (Phonologically Transformable Lexicon): A 35,000-word Computer Readable Pronouncing American English Lexicon on Structural Principles, with Accompanying Phonological Transformations, and Word Frequencies.* [Online]. Los Angeles: Copyright House Ear Institute. [Accessed].

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

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

Stacey, P. C., Raine, C. H., O'donoghue, G. M., Tapper, L., and Twomey, T. (2010). Effectiveness of computer-based auditory training for adult users of cochlear implants. *Int. J. Audiol.* 49, 347–356.

Stevens, K. N. (1998). *Acoustic Phonetics.* Cambridge, MA: MIT Press.

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

Summerfield, A. Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, eds B. Dodd and R. Campbell (London: Lawrence Erlbaum Associates, Inc.), 3–52.

Ungerleider, L. G., and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165.

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.

Vitevitch, M. S., and Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *J. Mem. Lang.* 40, 374–408.

Wang, M. D., and Bilger, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *J. Acoust. Soc. Am.* 54, 1248–1266.

Williams, P., and Simons, D. (2000). Detecting changes in novel, complex three-dimensional objects. *Vis. cogn.* 7, 297–322.

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Commun.* 26, 23–43.

Zeki, S. (2005). The Ferrier lecture 1995: behind the seen: the functional specialization of the brain in space and time. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1145–1183.

Zeng, F.-G., Popper, A. N., and Fay, R. R. (2004). *Cochlear Implants: Auditory Prostheses and Electrical Hearing*. New York: Springer.

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

Received: 04 December 2012; accepted: 28 February 2013; published online: 18 March 2013.

Citation: Bernstein LE, Auer ET, Eberhardt SP 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

This article was submitted to Frontiers in Auditory Cognitive Neuroscience, a specialty of Frontiers in Neuroscience.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on