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.,

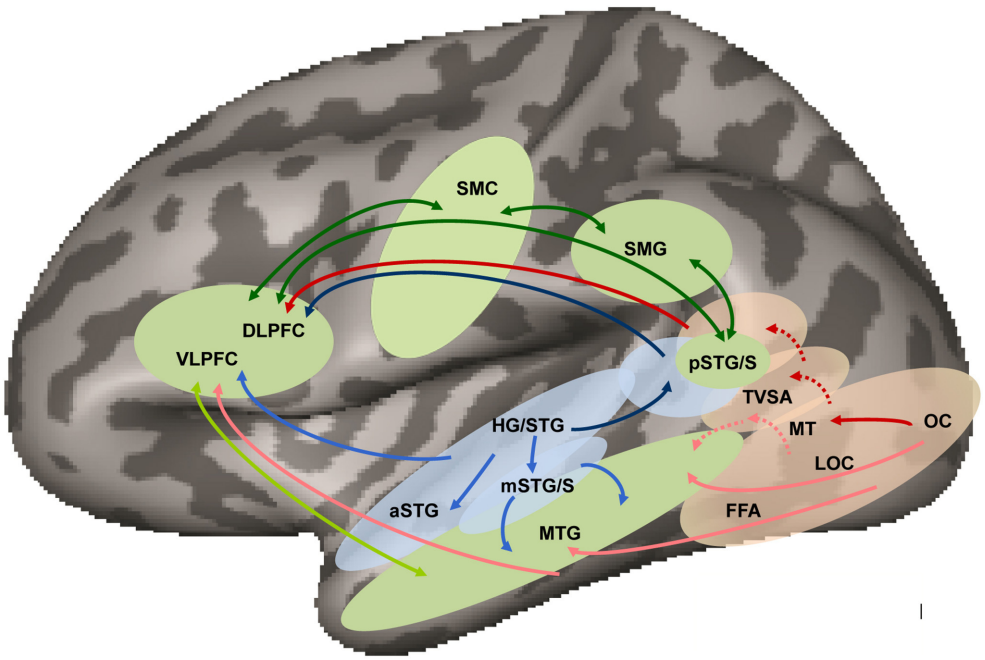

FIGURE 1 | Neuroanatomical working model of audiovisual speech perception in the left hemisphere based on models of dual visual (Wilson et al., 1993; Haxby et al., 1994; Ungerleider et al., 1998; Weiner and Grill-Spector, 2013) and auditory (Romanski et al., 1999; Hickok and Poeppel, 2007; Saur et al., 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 *sub-visemic* 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.