HHS Public Access

Author manuscript

Cortex. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

Cortex. 2015 July ; 68: 169–181. doi:10.1016/j.cortex.2015.03.006.

Prediction and constraint in audiovisual speech perception

Jonathan E. Peelle[1] and **Mitchell S. Sommers**[2]

[1]Department of Otolaryngology, Washington University in St. Louis, St. Louis MO USA

[2]Department of Psychology, Washington University in St. Louis, St. Louis MO USA

Abstract

During face-to-face conversational speech listeners must efficiently process a rapid and complex stream of multisensory information. Visual speech can serve as a critical complement to auditory information because it provides cues to both the timing of the incoming acoustic signal (the amplitude envelope, influencing attention and perceptual sensitivity) and its content (place and manner of articulation, constraining lexical selection). Here we review behavioral and neurophysiological evidence regarding listeners' use of visual speech information. Multisensory integration of audiovisual speech cues improves recognition accuracy, particularly for speech in noise. Even when speech is intelligible based solely on auditory information, adding visual information may reduce the cognitive demands placed on listeners through increasing precision of prediction. Electrophysiological studies demonstrate oscillatory cortical entrainment to speech in auditory cortex is enhanced when visual speech is present, increasing sensitivity to important acoustic cues. Neuroimaging studies also suggest increased activity in auditory cortex when congruent visual information is available, but additionally emphasize the involvement of heteromodal regions of posterior superior temporal sulcus as playing a role in integrative processing. We interpret these findings in a framework of temporally-focused lexical competition in which visual speech information affects auditory processing to increase sensitivity to auditory information through an early integration mechanism, and a late integration stage that incorporates specific information about a speaker's articulators to constrain the number of possible candidates in a spoken utterance. Ultimately it is words compatible with both auditory and visual information that most strongly determine successful speech perception during everyday listening. Thus, audiovisual speech perception is accomplished through multiple stages of integration, supported by distinct neuroanatomical mechanisms.

1. Introduction

Conversational speech can arrive at speeds exceeding 200 words per minute (Miller, Grosjean, & Lomanto, 1984). Speech comprehension must therefore be accomplished

Please address correspondence to: Dr. Jonathan Peelle, Department of Otolaryngology, Washington University in St. Louis, 660 South Euclid, Box 8115, St. Louis, MO 63110, peellej@ent.wustl.edu, 314-362-9044.

rapidly so that listeners can keep up with incoming information—a challenging task given the underarticulation, semantic ambiguity, acoustic variability, and background noise frequently present in everyday listening. To maximize efficiency during speech processing, listeners rely not only on the acoustic signal, but also on linguistic information (including lexical, syntactic, and semantic context) and additional sensory cues. Of chief importance among non-acoustic cues is the visual speech information available during face-to-face conversation. Here we review key types of information provided by visual speech, and the ways in which auditory and visual speech cues might be combined during natural listening. We broadly classify these effects into prediction (mechanisms which shape the processing of future sensory inputs) and constraint (information that aids in object identification by imposing restrictions on perceptual interpretations). Visual information affects speech processing in both of these ways, limiting the number of lexical interpretations and increasing the precision of listeners' predictions about the upcoming speech signal, facilitating processing if these predictions are realized (Gagnepain, Henson, & Davis, 2012; Sohoglu, Peelle, Carlyon, & Davis, 2012).

2. Information provided by visual speech

Visual speech refers to information available from seeing a speaker's mouth, including the lips, tongue, and teeth. Visual speech cues provide temporal markers corresponding to acoustic properties of a target speech signal, as well as specific information that helps resolve the identity of individual phonemes.

At the most basic level, articulatory movements provide listeners with an indication of when they should begin attending to a speaker. For example, in a noisy restaurant, knowing when our conversational partner is speaking is useful because it allows us to increase our attentional allocation to the target signal, aiding auditory stream segregation (Carlyon, Cusack, Foxton, & Robertson, 2001). In connected speech, mouth movements also help convey the temporal amplitude envelope of the speech, with an open mouth typically corresponding to a louder amplitude (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Tuning into rhythmic properties of connected speech may help listeners to know when to expect certain types of acoustic speech information and thus play a role in decoding syllabic and lexical categories (Peelle & Davis, 2012).

In addition to temporal information about the acoustic signal, visual speech contains cues regarding the position of a speaker's articulators that can complement acoustic speech information and that may be particularly important for speech understanding when auditory information is degraded. For example, place of articulation provides critical distinctions between certain consonants (e.g., /b/ vs. /d/ vs. /g/). This information is signaled acoustically by differences in the second formant (F2) frequency. Extracting place information from the acoustic signal, however, can be challenging in situations where F2 is susceptible to masking, including noisy environments and in listeners with impaired hearing. Under such conditions, the availability of place information in the visual signal provides a complementary source of information that can serve to distinguish whether, for instance, a speaker said “cat” or “cap”. Consequently, it is not surprising that speech perception is

almost always better when listeners can both see and hear a talker compared with listening alone (Erber, 1975; Sumby & Pollack, 1954).

In auditory speech, the phoneme is considered the unit of minimal distinction. The analog in visual speech is the viseme, referring to speech gestures that are confused during visual-only presentation (Fisher, 1968). A critical limitation of visual-only speech signals is that there is a many-to-one mapping between the basic units of auditory speech perception (phonemes) and the basic units of visual speech perception (visemes). That is, whereas /p/ and /b/ are readily distinguished acoustically by differences in voice onset times, visually they are nearly identical. Thus, although visual speech does not offer additional information compared to auditory-only speech for every phoneme, in many cases it can help disambiguate similar-sounding speech sounds. As we review in the next section, the result is a near ubiquitous improvement in speech perception for audiovisual speech relative to auditory-only speech.

3. Visual enhancement during speech perception

3.1 Performance benefits during audiovisual perception

The empirical benefits of audiovisual speech are frequently demonstrated during speech recognition in noise: Recognition for audiovisual speech in noise is substantially better than auditory speech alone, including improved recognition accuracy (Tye-Murray, Sommers, & Spehar, 2007a) and being able to perceive speech at a more favorable signal-to-noise ratio (SNR) (Bernstein, Auer Jr., & Takayanagi, 2004; Grant & Seitz, 2000; Macleod & Summerfield, 1987; Sumby & Pollack, 1954). However, advantages are also seen for audiovisual speech in measures that go beyond recognition. For example, when asked to shadow a spoken passage (i.e., repeat words in real time), listeners can perform the task more rapidly for audiovisual speech compared to auditory-only speech (Reisberg, McLean, & Goldfield, 1987). Audiovisual speech presentation results in better comprehension for short stories (Arnold & Hill, 2001) and in improved perceptual learning of degraded speech (Wayne & Johnsrude, 2012). Audiovisual presentations have also been found to reduce perceptual effort during speech-in-noise perception in the context of a dual-task paradigm (Gosselin & Gagné, 2011). Such findings are all consistent with the relatively uncontroversial stance that listeners make use of both visual and auditory information during speech perception, and that visual information aids recognition.

The behavioral advantage seen for audiovisual speech compared to auditory-only speech is frequently referred to as “visual enhancement”, implicitly reflecting both the fact that auditory information remains the primary cue to speech information, and that visual information is consistently seen as improving performance. Interestingly, however, not all listeners show the same benefit from visual speech information (Grant, Walden, & Seitz, 1998; Sommers, Tye-Murray, & Spehar, 2005; Tye-Murray, et al., 2007a). That is, individual listeners differ in the amount of visual enhancement they demonstrate, even after accounting for visual-only speech recognition (i.e., lipreading ability). The fact that listeners differ in the degree to which they make use of visual speech information has led to the suggestion that audiovisual integration is a discrete ability which may be preserved (or not) independently of auditory encoding and lipreading abilities (Grant, et al., 1998).

Conceptualizing multisensory integration as a discrete process can help explain individual differences in audiovisual speech perception that go beyond what would be predicted based on a straightforward combination of unimodal performance abilities.

3.2 Measurements of audiovisual integration

As noted above, assessing audiovisual integration independent from auditory-only or visual-only encoding is complicated by individual differences in unimodal encoding abilities. Two general classes of approach have been used to obtain estimates of individuals' ability to combine auditory and visual information that is independent of their ability to encode the unimodal stimuli. The first are modeling approaches in which predictions about audiovisual performance are obtained from measures of auditory and visual scores and deviations between predicted and observed audiovisual scores are attributed to differences in integration (Blamey, Cowan, Alcantara, Whitford, & Clark, 1989; Braida, 1991). A limitation of such approaches, however, is that models used to derive predicted audiovisual performance measures require assumptions about the integration process that can lead to conflicting results. For example, Braida (1991) proposed a model of audiovisual integration in which predicted audiovisual performance is based on an optimal observer model of auditory and visual integration. In this model participants typically obtain lower than predicted audiovisual scores. In contrast, Blamey (1989) proposed a simple probabilistic model of audiovisual integration in which individuals only make errors in an audiovisual condition if they make errors in both unimodal conditions. Under these assumptions, participants almost universally obtain higher-than-predicted audiovisual scores. In addition to relying on untested assumptions about the nature of integration, the modeling approach to audiovisual integration requires confusion matrices for the two unimodal conditions in order to obtain measures of predicted audiovisual performance. Consequently, this approach has been restricted to consonant-vowel stimuli in which obtaining the required consonant confusion matrices is relatively straightforward (but also quite time-consuming).

The second approach that has been used to assess audiovisual integration is to measure relative performance on tasks that necessarily require audiovisual integration, such as the McGurk effect (McGurk & MacDonald, 1976). In this well-known audiovisual illusion, listeners see and hear discrepant auditory and visual information (e.g., hearing /ba/ while simultaneously watching someone mouth /ga/), with listeners typically reporting a fused percept (for the case above either /da/ or /tha/). However, susceptibility to the McGurk effect depends on factors other than audiovisual integration, including differences in lipreading and the ability to detect incongruity between the auditory and visual signals.

A second task-based measure that has been used to assess integration is asynchrony detection (Grant & Seitz, 1998). Thresholds for asynchrony detection are typically measured by presenting auditory and visual stimuli that are temporally offset and measuring thresholds for detecting the temporal asynchrony. The rationale behind this approach is that poorer integrators are expected to have higher asynchrony detection thresholds. However, Grant and Seitz (1998) found no correlation between asynchrony detection and speechreading abilities. Thus, the degree to which crossmodal synchrony detection relates to audiovisual speech perception remains unclear.

In summary, although there is general agreement that listeners use visual information during audiovisual speech processing, there is less consensus on the mechanisms through which visual information acts to influence recognition and perception. This issue is particularly interesting given that visual enhancement varies as a function of acoustic clarity and linguistic content (Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014), suggesting that audiovisual integration can be tailored to specific listening situations. Given the multiple cues provided by visual speech, it would be surprising if only a single dimension was useful for listeners. Here we focus on two complementary aspects of visual speech: temporal information provided by the mouth opening, and constraints on phoneme identity provided by specific articulator position.

3.3 Visual speech aids temporal prediction

Because visual speech conveys information about the amplitude envelope of the acoustic signal, it may help listeners know when to attend to an auditory signal. Temporally allocating attention is critical in perceptual processing, particularly for faint or near-threshold stimuli. In a wide variety of non-speech paradigms, temporal expectancy can significantly modulate the detectability of sensory information. For example, if an auditory tone is presented at a predictable time, participants are better able to detect its presence in noise compared to an irregular interval (Egan, Greenberg, & Schulman, 1961; Watson & Nichols, 1976).[1]

Attentional enhancement of sensory detection also operates cross-modally: visual cues can enhance perception and allow listeners to detect tones at a softer level. For example, ten Oever and colleagues (2014) presented tones to participants that occurred at regular or irregular intervals. In some cases the tones were preceded by a brief visual cue. Consistent with previous studies, the authors found that auditory detection thresholds were lower (i.e., listeners were more sensitive) for rhythmic compared to random presentations. Critically, however, detection thresholds were lower during audiovisual compared to auditory-only presentation. Conceptually similar findings were reported by Tye-Murray et al. (2011) who compared detection thresholds for auditory-only presentation of the syllable /ba/ with three different audiovisual conditions. The auditory signal was identical to the one presented in the auditory-only condition; visual information consisted of an unaltered video clip of a talker's face, a low-contrast version of the same clip, or a mouth-like Lissajous figure. Although the benefits of visual information for detection thresholds varied across the three different visual conditions, performance was better for all of the audiovisual conditions than for the auditory-only condition. These findings are consistent with crossmodal influences on temporal attention that increase sensitivity to expected sensory information.

How might such visual influences on temporal attention translate to connected speech? Grant and Seitz (2000) investigated the impact of visual speech information on detectability for spoken sentences in noise. They presented participants with auditory-only sentences, auditory sentences with visual information from the same sentence (matching), or auditory sentences coupled with visual information from a different sentence (mismatching). If the

[1]Participants in Watson and Nichols (1976) may have been particularly motivated, as on a subset of trials they received an electric shock to their ankle if they failed to respond appropriately.

correspondence between auditory and visual information aids detection, detection should be best in the matching condition. Indeed, average detection thresholds were lower when the auditory and visual presentations were matched. Further analysis suggested that the audiovisual benefit obtained was related to the correlation between mouth opening and the amplitude envelope of the acoustic speech signal. The shift in detection threshold demonstrates that facilitation provided by congruent audiovisual information is not restricted to word recognition, but extends to perceptual sensitivity.

Knowing when a perceptual event is likely to occur can help detection. But how does visual speech information help listeners identify the actual words they are hearing? In the next section we introduce a framework in which visual speech information constrains the number of candidate words associated with a particular speech token.

3.4 Visual speech constrains lexical competition

When listening to speech, our goal is to correctly identify the words we are hearing so that we can extract the intended meaning. This process of identification can be understood within a framework of lexical competition in which a spoken input activates a set of phonologically similar candidate items and the correct target must be chosen from among that set of competitors, or neighbors (Luce & Pisoni, 1998; Marslen-Wilson & Tyler, 1980).[2] In the domain of spoken language, phonological (auditory) neighbors are frequently operationalized as words that differ by the addition, deletion, or substitution of a single phoneme with a target word, as illustrated in Figure 1. The set of competitor items is frequently referred to as the phonological neighborhood of a target word, with the number of neighbors referred to as the neighborhood density. Target words with high neighborhood density are typically more difficult to recognize because they receive more competition from similar-sounding words. That is, the activation of lexical candidates is relatively imprecise because of overlapping perceptual features, making a single target word difficult to select.

Visual speech information also lends itself to a neighborhood framework in which the viseme functions as the basic unit. Visual-only lexical competition is thus constrained by the visual similarity of speech gestures in an analogous way as auditory speech is constrained by acoustic similarity and, as with auditory neighborhoods, visual neighborhood density is inversely correlated with performance in visual-only conditions (Auer Jr., 2002; Feld & Sommers, 2011; Mattys, Bernstein, & Auer Jr., 2002).

Tye-Murray, Sommers, & Spehar (2007b) extended the activation-competition framework that has been successful as a model of unimodal speech perception to audiovisual presentations. The authors suggested that audiovisual speech recognition was generally better than either of the unimodal conditions because competition in the bimodal case is restricted only to candidate items that are neighbors in both auditory-only and visual-only modalities (i.e., the intersection density). Importantly, across different words, the intersection density can dissociate from auditory and visual neighborhood size. Thus, words

[2]Although in many situations the onset of a word may have special importance in the competition process, we use the term “competitor” or “neighbor” to refer to all words that may be empirically confused with a target word, including those whose onsets differ.

with similar auditory-only and visual-only neighborhoods can still differ in intersection density, as shown in the bottom portion of Figure 1. A critical question arises concerning how auditory and visual constraints are considered during speech comprehension. Does lexical competition scale with auditory neighborhood density, visual neighborhood density, or intersection density?

To answer this question, Tye-Murray and colleagues (2007b) tested adult listeners using a repetition task in which target words spoken in 6-talker babble were presented in auditory-only, visual-only, and audiovisual speech conditions. As expected, participants' ability to accurately repeat words was significantly higher in the audiovisual condition than in either unimodal condition. Critically, performance differences in the audiovisual speech condition were attributed to intersection density, with words having a larger intersection density resulting in poorer recognition than words with a smaller intersection density. Neither the auditory-only nor the visual-only neighborhood density related to audiovisual performance. These findings suggest that lexical constraints provided by both auditory and visual information jointly act to constrain participants' recognition of speech.

4. Models of multisensory integration

Several frameworks have been proposed to describe how multisensory information is combined during speech perception. Three classes of models are shown in Figure 2, and reflect the following shared assumptions:

1. Auditory and visual information from the speech signal are necessarily processed “separately” due to the physiological dissociation between auditory and visual speech information in sensory cortices.
2. At some point prior to lexical identification auditory and visual information are integrated.
3. Lexical information acts to constrain perception.

An influential model of audiovisual speech perception was proposed by Grant and colleagues (1998). Shown in Figure 2a, in this view auditory and visual information are processed separately to begin with and combined later in a separate integration stage. Following audiovisual integration, lexical identity is determined, which can then feed back to auditory and visual processing separately. We refer to this as a “late integration” model because audiovisual integration occurs only after unimodal processing has completed. An advantage of late integration models is that they partition the influence of integration into a discrete cognitive stage, which may help account for individual differences in audiovisual speech perception that cannot be predicted by unimodal performance. Late integration theories may also provide a mechanism through which listeners can differentially weigh auditory and visual information depending on how informative each modality is.

An alternative view is that integration occurs earlier on—that is, that processing within one modality can be influenced by another (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). This view is shown in Figure 2b (illustrated to emphasize the dominance of auditory information, although in theory “early integration” might also be bidirectional). Early

integration models are motivated in part by electrophysiological data showing multimodal responses in primary sensory cortices (Schroeder & Foxe, 2005). Early integration models provide a mechanism for visual information to increase perceptual sensitivity to auditory signals, and receive strong support from electrophysiological recordings made in auditory cortex (covered in more detail below).

Finally, we suggest a third, hybrid family of models that incorporate integration at multiple stages (and through multiple mechanisms). Figure 2c depicts a hybrid model that straightforwardly combines the early and late integration views discussed above. Thus, visual cues can act to shape auditory perception (early integration, via auditory cortex), but may also be incorporated at a later stage (anatomically distinct from auditory cortex). Multistage integration allows for both early influences of visual cues on auditory processing, and a later stage of integration that combines outputs of lower-level processing (including differential weighing of modalities based on their reliability). Multistage models may explain in part the difficulty of predicting audiovisual speech performance based on a single measure of integration, as multiple integration mechanisms may interact in ways not captured by single-stage models.

Evidence regarding the utility of these various frameworks comes from many sources. For example, if audiovisual integration is a single, discrete processing stage, individual differences in integration ability should be measurable and relate to multisensory performance on a variety of tasks (but see above for limitations of task-based approaches to integration). Alternatively, neurophysiological evidence for either multiple cortical regions involved in integration, or for different physiological processes, would be more consistent with multistage integration than single-stage integration (although one could, in theory, suggest a single conceptual stage that is carried out by multiple neural processes).

5. Neural mechanisms supporting audiovisual speech processing

Understanding how auditory and visual information are combined is critical in differentiating models of multisensory integration illustrated in Figure 2. Intuitively, it seems reasonable to assume that auditory and visual speech information are processed predominantly in their respective sensory cortices. However, subcortical pathways and cortico-cortico connections leave open the possibility for crossmodal interactions in early cortical areas. Neuroscientific approaches to understanding audiovisual speech processing have provided useful insights to these issues in the context of both temporal prediction and the potential anatomical structures responsible for multisensory integration.

Whereas behavioral measures of audiovisual speech perception typically focus on increased accuracy in audiovisual conditions compared to auditory-only conditions (visual enhancement), neuroimaging studies can potentially access a more direct measure of multisensory integration by measuring the neural responses to audiovisual speech compared to auditory-only or visual-only presentations. Several views on how to assess multimodal integration have been proposed (reviewed in Stevenson et al., 2014), the most common of which are illustrated in Figure 3. If the response to audiovisual speech is identical to that for auditory-only speech, it is commonly assumed that activity in a region does not reflect

integration. Somewhat paradoxically, if audiovisual speech results in any difference in activity compared to auditory only—that is, either an increase or a decrease—this is frequently interpreted as reflecting multisensory integration (although this is still a point of debate).

Neurophysiological evidence may be particularly relevant to the question of whether multisensory integration happens at an early or a late stage (Braida, 1991; Grant, et al., 1998; Massaro, 1999; Oden & Massaro, 1978). That is, does audiovisual integration occur after unimodal inputs have been fully processed, or are multisensory effects integrated into online auditory processing? Anatomically, these positions can be distinguished by assessing whether visual information impacts processing in auditory cortex, or higher-level heteromodal association regions. Below we review neuroimaging evidence that speaks to how visual speech information aids listeners' prediction of the timing and identity of spoken language.

5.1 Oscillations in auditory cortex track the speech amplitude envelope

As recorded from scalp EEG, MEG, or local field potentials obtained from electrodes, oscillations in cortical activity reflect synchronized fluctuations in membrane potential across a large population of neurons. These systematic variations in cellular excitation mean that inputs to a cell arriving at some phases are more likely to trigger a response than those arriving at other phases, as illustrated in Figure 4. Thus, near-threshold stimuli presented at a high-excitability phase of oscillation are more likely to be perceived than those arriving at a low-excitability phase. This phenomenon is apparent in both visual (Romei, Gross, & Thut, 2010; Thut, Nietzel, Brandt, & Pascual-Leone, 2006) and auditory (Lakatos et al., 2005) modalities. An important property of neural oscillations is that activity at different frequencies can be nested: that is, the phase of low frequency oscillations (e.g., theta: ∼4–8 Hz) can modulate power at higher frequency oscillations (e.g., gamma: 30+ Hz) (Canolty et al., 2006; Lakatos, et al., 2005). This nested hierarchy of cortical oscillations provides a mechanistic framework for coordinating neural activity on different timescales, and across multiple brain regions, with both each other and the sensory input (Canolty & Knight, 2010; Fries, 2005; Jensen & Colgin, 2007).

Ongoing low-frequency (< 8 Hz) neural oscillations are increasingly seen as playing an important role in how listeners process connected speech (Giraud & Poeppel, 2012; Peelle & Davis, 2012). Cortical oscillations in this frequency range show phase locking to unintelligible auditory stimuli that are enhanced when speech has sufficient acoustic detail to become intelligible (Doelling, Arnal, Ghitza, & Poeppel, 2014; Luo & Poeppel, 2007; Peelle, et al., 2013). Oscillatory phase reflects the excitability of neural populations, and as such phase-locked oscillatory responses impact the ease with which a sensory signal can be processed: detection thresholds and response time vary as a function of oscillatory phase (Henry, Herrmann, & Obleser, 2014; Henry & Obleser, 2012; Lakatos, et al., 2008; Lakatos, et al., 2005). Phase-locked oscillatory activity also tracks shifts in attention, and during multi-talker environments reflects the attended-to speech stream (Kerlin, Shahin, & Miller, 2010; Mesgarani & Chang, 2012; Zion Golumbic et al., 2013). These phase-locked cortical responses appear to track the acoustic amplitude envelope of the speech signal,

corresponding approximately to syllable rate in connected speech. Although many empirical findings have focused on these syllable-rate oscillations, such entrainment has the potential to impact the neural processing of speech on multiple timescales through cross-frequency coupling (Ghitza, 2011; Giraud & Poeppel, 2012; Gross et al., 2013).

The fact that visual speech provides clues to the acoustic amplitude envelope suggests that visual information might help modulate online tracking of the acoustic speech signal (Schroeder, et al., 2008). Evidence from animal models indicates that primary auditory regions are in fact multisensory (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007; Schroeder & Foxe, 2005). Critically, electrophysiological recordings in nonhuman primates have demonstrated that non-auditory stimuli can reset the phase of low-frequency oscillations in auditory cortex (Kayser, Petkov, & Logothetis, 2008; Lakatos, et al., 2007; Perrodin, Kayser, Logothetis, & Petkov, 2015). In this context, we might expect that quasi-rhythmic visual activity associated with the movements of articulators and mouth opening would be reflected in phase-locked cortical activity in auditory cortex during audiovisual speech perception.

To test this hypothesis, Luo et al. (2010) presented participants with movies containing both auditory and visual information that were matching or mismatching, allowing the authors to investigate the degree to which early auditory regions tracked auditory information, visual information, and their interaction. The authors indeed found more consistent neural phase relationships in auditory cortex in response to the same acoustic input when the visual input matched the auditory signal compared to when it did not. These results suggest that congruent visual information can increase the precision of oscillatory auditory activity in the 4–8 Hz range (for an example with gesture, see Biau, Torralba, Fuentemilla, de Diego Balaguer, & Soto-Faraco, In press).

Enhancing phase-locked oscillatory responses is of interest to the degree that it aids listeners in speech comprehension. In auditory-only speech, attention has been shown to increase the envelope-tracking response to an attended speaker compared to an ignored speaker (Ding & Simon, 2012; Kerlin, et al., 2010; Mesgarani & Chang, 2012; Zion Golumbic, Ding, et al., 2013). A ubiquitous problem in speech perception is understanding a target speaker in the presence of background noise or a competing talker. In this “cocktail party” environment, the ability to direct attention to the target speech signal is especially critical.Zion Golumbic and colleagues (2013) presented listeners with movies containing visual information and either a single (audiovisual matching) speaker, or two competing speakers, with instructions to attend to one of the voices. Participants were also presented with auditory-only analogs of the same conditions. The authors found that the amplitude of the oscillatory entrainment to the attended talker was increased in the audiovisual condition compared to auditory-only speech. The increased responses were localized to auditory cortex and not observed elsewhere (including visual cortex).

Together, these results are consistent with the hypothesis that visual speech information increases the accuracy with which oscillations in auditory cortex track the ongoing speech signal, improving speech perception accuracy.

5.2 Multisensory effects in auditory and posterior superior temporal cortices

A great deal of neuroimaging research on audiovisual integration has focused on the posterior portion of the superior temporal sulcus (STS). The posterior STS receives inputs from both auditory and extrastriate visual cortices (Seltzer & Pandya, 1978), and thus seems anatomically well-suited to perform multisensory integration. In fMRI studies, posterior STS responds to both visual and auditory input (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004), and in many cases shows supra-additive responses to audiovisual speech (Calvert, Campbell, & Brammer, 2000; Sekiyama, Kanno, Miura, & Sugita, 2003; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003).

It is worth noting that many studies of audiovisual speech processing rely on syllable stimuli in which auditory and visual information are altered in their level of congruency (i.e., McGurk stimuli). However, although syllabic stimuli contain both auditory and visual speech information, they lack the lexical constraints present in conversational listening. In addition, metalinguistic categorization tasks may not reflect the type of executive or attentional demands used during everyday listening. Thus, it is important to also consider neural responses to audiovisual speech that includes words or phrases. In at least one such study, posterior STS has also been shown to be preferentially active for audiovisual sentences compared to auditory-only sentences (Yi, Smiljanic, & Chandrasekaran, 2014).

In an fMRI study of audiovisual speech processing, Nath and Beauchamp (2011) varied whether auditory or visual information was more reliable by varying the unimodal perceptual clarity of audiovisual syllables (that is, presenting a clear auditory signal with a blurred visual signal, or a clear visual signal with an auditory signal in noise). Although posterior STS showed a similar overall magnitude of response for various audiovisual speech conditions, the functional connectivity of the STS was modulated as a function of the sensory clarity in each modality: functional connectivity was greater from STS to auditory cortex when auditory information was clearer, and relatively greater to visual cortex when visual information was clearer. These results are consistent with an online weighting of auditory and visual information in which listeners dynamically adjust to stimulus clarity. They also suggest an alternate explanation for the involvement of posterior STS in audiovisual speech perception. The frequent finding of an increased response of posterior STS during audiovisual speech relative to auditory-only speech certainly suggests a role in multisensory processing, and is often assumed to reflect multisensory integration. However, it may also be that posterior STS is active in determining relative weighting or attention directed at complementary modalities. Consistent with this view, activity in posterior STS also appears to distinguish between variations in visual clarity to a greater degree when speech is less intelligible (McGettigan et al., 2012).[3]

Multisensory effects are also observed in early auditory areas. For example, audiovisual speech (which should be a more predictable stimulus than auditory-only speech) results in increased activity in primary and secondary auditory cortices (Okada, Venezia, Matchin,

[3]Given that audiovisual integration processes may vary as a function of acoustic clarity, it is worth considering that some of the variability in results of fMRI studies may be attributable to different levels of acoustic challenge (e.g., sparse vs. continuous scanning) (Peelle, 2014).

Saberi, & Hickok, 2013), consistent with increased activity in auditory cortex during silent lipreading (Calvert et al., 1997). Audiovisual interactions in auditory cortex appear to precede those in posterior STS (Möttönen, Schürmann, & Sams, 2004). Together these observations indicate that visual information impacts auditory processing at the earliest stages of the cortical hierarchy. Thus, although there is strong evidence that posterior STS plays a role in audiovisual speech perception, its role must be interpreted in the context of primary auditory cortex showing multisensory responses.

Finally, it is worth noting that several researchers have also implicated motor regions in audiovisual speech processing (Fridriksson et al., 2008; Hall, Fussell, & Summerfield, 2005; Skipper, Nusbaum, & Small, 2005). One interpretation of these findings is that observing a visual speech gesture may activate motor plans, which shape auditory perceptual experience (Tian & Poeppel, 2012). That is, observing a speaker's articulators causes motor activity in a listener that in turn preferentially activates phonemic categories (Möttönen & Watkins, 2009). One recent finding consistent with the role of motor plans in visual speech perception is the self-advantage for lipreading: After controlling for individual differences in lipreading ability and differences in how well a particular speaker can be lipread, lipreading scores are higher when individuals viewed themselves as the target talker compared to when others served as the speaker (Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2013, In press). Thus, although several suggestive lines of evidence implicate the motor system in audiovisual speech perception, there is not a clear consensus. It is likely that both the level of linguistic processing (Peelle, 2012), the stimulus clarity (McGettigan, et al., 2012), and the congruence of auditory and visual speech information (Erickson, Heeg, Rauschecker, & Turkeltaub, 2014) play important roles here. On balance, converging functional and anatomical evidence point most strongly towards auditory cortex and posterior STS as playing the most prominent roles in multisensory integration during speech perception. Such an arrangement is most consistent with a multistage integration process.

6. Prediction and constraint in speech perception

As we have seen, it is clear from behavioral studies that listeners make use of visual speech information to aid speech perception. Given that articulator movement for a phoneme can precede acoustic information by 100–200 ms (Chandrasekaran, et al., 2009), the temporal order of both the sensory input and electrophysiological effects suggests that visual speech information may provide predictions about upcoming auditory input.[4] On the other hand, participants will sometimes perceive crossmodal events jittered by up to 250 ms as occurring simultaneously (with an asymmetry favoring an auditory lag; van Wassenhove, Grant, & Poeppel, 2007), suggesting that physical timing is not the only determining factor in multisensory integration. To what degree can visual speech information be considered predictive?

From one perspective, neural oscillations by definition encode predictions (Arnal & Giraud, 2012; Engel, Fries, & Singer, 2001). That is, the phase of low-frequency oscillations

[4]Important caveats to these points are found in Schwartz and Savariaux (2014), who argue that in connected speech the range of visual-auditory correspondence can vary widely, with consistent visual leads being most apparent only in preparatory gestures. However, they also note that visual leading is not necessary for visual information to aid in the prediction of auditory speech.

determines perceptual sensitivity; when stimuli are sufficiently regular, oscillatory phase corresponds to the likely occurrence of the next event. Events that correspond to the prediction (that is, occur during a high-excitability portion of the phase) are processed more rapidly and efficiently than events which do not (Henry & Obleser, 2012; Lakatos, et al., 2008). Thus, to the extent that non-auditory speech information acts to reset the phase of low-frequency neural oscillations, visual speech necessarily acts to aid in prediction.

Additional evidence in support of prediction would come from situations in which congruent audiovisual speech cues would not only aid recognition, but detection. That is, visual speech would affect processing on a perceptual level, not merely at a post-perceptual stage. Behavioral evidence from audiovisual sentence processing supports this view, with audiovisual speech associated with improved detection thresholds relative to auditory-only speech (Grant & Seitz, 2000; Tye-Murray, et al., 2011).

Neural evidence regarding prediction can be interpreted in a predictive coding framework in which error signals reflect a mismatch between predictions and sensory information. In the most straightforward view of predictive coding, to the degree that predicted sensory inputs are typically processed more easily (Friston, 2010; Friston & Kiebel, 2009), if visual speech cues are indeed predictive then audiovisual speech should result in reduced neural processing relative to unimodal speech input.

If visual speech indeed provides predictions that listeners use to parse upcoming auditory signals, we would expect that the degree of facilitation would vary with the amount of predictive information provided by a visual cue. This cue-specific facilitation is precisely what was found by van Wassenhove et al. (2005) using a McGurk paradigm: temporal facilitation of time-locked EEG signals depended on how accurate participants were on visual-only identification. That is, the least neural facilitation was present for /ka/ (which participants correctly identified ∼65% of the time using visual information alone), and the most facilitation for /pa/ (where visual-only identification was over 95%). (It should be noted that in the same study, audiovisual speech resulted in overall lower amplitude ERPs, regardless of congruency, suggesting a complex interaction of multisensory integration and congruency in generating the timing and magnitude of neural responses.)

Within a predictive coding framework, the converse of congruency-related facilitation is that incongruent stimuli—where auditory information does not match visually-generated predictions—should generate an error signal. Arnal and colleagues (2011) explicitly tested this hypothesis in the context of a McGurk paradigm. They found that incongruent audiovisual speech stimuli elicited a larger response beginning around 300 ms after speech onset. Importantly, and consistent with the facilitation results reported by van Wassenhove et al., the degree of error was related to the strength of the prediction provided by the visual-only stimuli, such that visemes that participants could more accurately identify based on visual information alone generated stronger error signals when the acoustic information failed to match these predictions.

However, although a straightforward predictive coding account is intuitively appealing, the story may be somewhat more complex. As noted above, congruent audiovisual speech

(which should be a more predictable stimulus than auditory-only speech) has been found to result in increased activity in primary and secondary auditory cortices (Okada, et al., 2013). A more complete understanding of audiovisual speech in the context of predictive coding will likely have to take a fuller account of the type of linguistic stimuli used, perceptual clarity, and type of responses required. It is also likely that specific task demands and attentional allocation will impact the result.

Together, the impact of visual speech information within a short time of acoustic onset in many of these studies—that is, affecting the time-locked neural responses within hundreds of milliseconds—suggests that visual information constrains speech perception in an online manner.

7. Multistage integration and lexical competition

Here we have proposed that visual information informs speech perception through multiple, complementary mechanisms. The first is through early influences on auditory cortex that shape perception in real-time—that is, visual input alters the processing of auditory information as it is being heard. The strongest evidence for early integration comes from crossmodal reset of low-frequency neural oscillations in auditory cortex. Increasing perceptual sensitivity through oscillatory entrainment provides listeners with more acoustic detail, and thus reduces lexical competition in much the same way that reducing the level of background noise would.

Complementary information about speech gestures (such as place or manner of articulation) also acts to improve perception. In some instances visual cues precede auditory processing and may shape early prediction; in other cases the visual cues may be processed in parallel. We argue that cues related to speech gestures are generally better thought of as affecting late integration. This is most clearly demonstrated in situations such as in the McGurk effect in which there is a conflict between auditory and visual information—in such situations a fused percept arises from incompatible unimodal input. Such an outcome strongly suggests the post-perceptual combining of information, for which posterior STS is well suited. Indeed, posterior STS is likely to play a role in weighting auditory and visual inputs (Nath & Beauchamp, 2011), generating predictions that are passed to sensory cortex (Arnal, et al., 2011).

The parallel influence of multisensory information during speech processing may explain some of the variability in experimental results. In this context, it is important to note that we view these multistage integration processes as being dynamic—the degree to which they operate will depend on the specific information available. For example, in an analogous way to auditory-only speech, it is likely that ongoing oscillatory entrainment plays significantly less of a role for single word perception in which no rhythmic sensory information is present compared to what occurs during sentence comprehension (Peelle & Davis, 2012). When processing isolated speech tokens it may be that late integration plays a more dominant role, and that auditory cortex is in a continuous processing mode that is less biased towards particular temporal windows (Lakatos, et al., 2008).

7.1 How specific is multisensory integration to speech?

We have focused on the processes supporting multisensory integration in the context of audiovisual speech perception. Of course, speech perception is only one situation in which human observers integrate information from multiple sensory modalities. To what degree are these integration processes unique to spoken language? Many of the brain regions discussed here have also been implicated in nonspeech paradigms. As noted above, electrophysiological studies are replete with examples of crossmodal influences of non-auditory stimuli on auditory cortical responses. In human fMRI studies, posterior STS shows an increased response to audiovisual nonspeech stimuli in a manner comparable to that seen for speech stimuli (Stevenson, Geoghegan, & James, 2007). Thus, on balance, it is likely that the neural processes that listeners rely upon during audiovisual speech perception are not necessarily unique to speech, and instead serve multisensory integration more generally. However, there may be aspects of human speech that differ from nonspeech stimuli due to their sensory complexity, semantic content, or listeners' extensive experience with these stimuli that impact processing (Vatakis, Ghazanfar, & Spence, 2008).

8. Conclusions

Speech comprehension is frequently a multisensory experience. Here we have reviewed evidence that visual information aids in predicting both the timing of acoustic speech events (primarily through the amplitude envelope) and the set of possible lexical candidates (primarily through position of the articulators). Accumulating neuroimaging evidence suggests that both mechanisms exert influence on auditory cortical areas, consistent with multistage integration processes involving both sensory and heteromodal cortices. These data speak to the coordinated influence of both early and late integration mechanisms operating during the normal processing of audiovisual speech.

Acknowledgments

This work was supported by The Dana Foundation and NIH grants R01AG038490 and R01AG018029. We are grateful to Avanti Dey and Kristin Van Engen for helpful comments on this work.

References

Arnal LH, Giraud AL. Cortical oscillations and sensory predictions. Trends inCognitive Sciences. 2012; 16:390–398.

Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature Neuroscience. 2011; 14:797–801.

Arnold P, Hill F. Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. British Journal of Audiology. 2001; 92:339–355.

Auer ET Jr. The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. Psychonomic Bulletin and Review. 2002; 9:341–347. [PubMed: 12120798]

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nature Neuroscience. 2004; 7:1190–1192.

Bernstein LE, Auer ET Jr, Takayanagi S. Auditory speech detection in noise enhanced by lipreading. Speech Communication. 2004; 44:5–18.

Biau E, Torralba M, Fuentemilla L, de Diego Balaguer R, Soto-Faraco S. Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. Cortex. In press.

Blamey P, Cowan R, Alcantara J, Whitford L, Clark G. Speech perception using combinations of auditory, visual, and tactile information. Journal of Rehabilitation Research and Development. 1989; 26:15–24. [PubMed: 2521904]

Braida LD. Crossmodal integration in the identification of consonant segments. Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology. 1991; 43:647–677.

Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SCR, McGuire PK, David AS. Activation of auditory cortex during silent lipreading. Science. 1997; 276:593–596. [PubMed: 9110978]

Calvert GA, Campbell R, Brammer MJ. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology. 2000; 10:649–657. [PubMed: 10837246]

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Kirsch HE, Knight RT. High gamma power is phase-locked to theta oscillations in human neocortex. Science. 2006; 313:1626–1628. [PubMed: 16973878]

Canolty RT, Knight RT. The functional role of cross-frequency coupling. Trends in Cognitive Sciences. 2010; 14:506–515. [PubMed: 20932795]

Carlyon RP, Cusack R, Foxton JM, Robertson IH. Effects of attention and unilateral neglect on auditory stream segregation. Journal of Experimental Psychology: Human Perception and Performance. 2001; 27:115–127. [PubMed: 11248927]

Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. PLoS Computational Biology. 2009; 5:e1000436. [PubMed: 19609344]

Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. Journal of Neurophysiology. 2012; 107:78–89. [PubMed: 21975452]

Doelling K, Arnal LH, Ghitza O, Poeppel D. Acoustic landmarks drive delta– theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage. 2014; 85:761–768. [PubMed: 23791839]

Egan JP, Greenberg GZ, Schulman AI. Interval of time uncertainty in auditory detection. Journal of the Acoustical Society of America. 1961; 33:771–778.

Engel AK, Fries P, Singer W. Dynamic predictions: Oscillations and synchrony in top-down processing. Nature Reviews Neuroscience. 2001; 2:704–716.

Erber NP. Auditory-visual perception of speech. Journal of Speech and Hearing Disorders. 1975; 40:481–492. [PubMed: 1234963]

Erickson LC, Heeg E, Rauschecker JP, Turkeltaub PE. An ALE meta-analysis on the audiovisual integration of speech signals. Human Brain Mapping. 2014; 35:5587–5605. [PubMed: 24996043]

Feld J, Sommers MS. There goes the neighborhood: Lipreading and the structure of the mental lexicon. Speech Communication. 2011; 53:220–228. [PubMed: 21170172]

Fisher CG. Confusions among visually perceived consonants. Journal of Speech and Hearing Research. 1968; 11:796–804. [PubMed: 5719234]

Fridriksson J, Moss J, Davis B, Baylis GC, Bonilha L, Rorden C. Motor speech perception modulates the cortical language areas. NeuroImage. 2008; 41:605–613. [PubMed: 18396063]

Fries P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in Cognitive Sciences. 2005; 9:474–480. [PubMed: 16150631]

Friston K. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience. 2010; 11:127–138.

Friston K, Kiebel S. Predictive coding under the free-energy principle. Philosophical Transactions of The Royal Society B. 2009; 364:1211–1221.

Gagnepain P, Henson RN, Davis MH. Temporal predictive codes for spoken words in auditory cortex. Current Biology. 2012; 22:615–621. [PubMed: 22425155]

Ghitza O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Frontiers in Psychology. 2011; 2:130. [PubMed: 21743809]

Giraud AL, Poeppel D. Cortical oscillations and speech processing: Emerging computational principles and operations. Nature Neuroscience. 2012; 15:511–517.

Gosselin PA, Gagné JP. Older adults expend more listening effort than younger adults recognizing audiovisual speech in noise. International Journal of Audiology. 2011; 50:786–792. [PubMed: 21916790]

Grant KW, Seitz PF. The use of visible speech cues for improving auditory detection of spoken sentences. Journal of the Acoustical Society of America. 2000; 108:1197–1208. [PubMed: 11008820]

Grant KW, Seitz PF. Measures of auditory-visual integration in nonsense syllables and senteences. Journal of the Acoustical Society of America. 1998; 104:2438–2450. [PubMed: 10491705]

Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. Journal of the Acoustical Society of America. 1998; 103:2677–2690. [PubMed: 9604361]

Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLOS Biology. 2013; 11:e1001752. [PubMed: 24391472]

Hall DA, Fussell C, Summerfield AQ. Reading fluent speech from talking faces: Typical brain networks and individual differences. Journal of Cognitive Neuroscience. 2005; 17:939–953. [PubMed: 15969911]

Henry MJ, Herrmann B, Obleser J. Entrained neural oscillations in multiple frequency bands comodulate behavior. Proceedings of the National Academy of Sciences. 2014; 111:14935–14940.

Henry MJ, Obleser J. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proceedings of the National Academy of Science. 2012; 109:20095–20100.

Jensen O, Colgin LL. Cross-frequency coupling between neuronal oscillations. Trends in Cognitive Sciences. 2007; 11:267–269. [PubMed: 17548233]

Kayser C, Petkov CI, Logothetis NK. Visual modulation of neurons in auditory cortex. Cerebral Cortex. 2008; 18:1560–1574. [PubMed: 18180245]

Kerlin JR, Shahin AJ, Miller LM. Attentional gain control of ongoing cortical speech representations in a “cocktail party”. Journal of Neuroscience. 2010; 30:620–628. [PubMed: 20071526]

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE. Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron. 2007; 53:279–292. [PubMed: 17224408]

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science. 2008; 320:110–113. [PubMed: 18388295]

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. Journal of Neurophysiology. 2005; 94:1904–1911. [PubMed: 15901760]

Luce PA, Pisoni DB. Recognizing spoken words: The neighborhood activation model. Ear and Hearing. 1998; 19:1–36. [PubMed: 9504270]

Luo H, Liu Z, Poeppel D. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biology. 2010; 8:e1000445. [PubMed: 20711473]

Luo H, Poeppel D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron. 2007; 54:1001–1010. [PubMed: 17582338]

Macleod A, Summerfield Q. Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology. 1987; 21:131–141. [PubMed: 3594015]

Marslen-Wilson WD, Tyler L. The temporal structure of spoken language processing. Cognition. 1980; 8:1–71. [PubMed: 7363578]

Massaro DW. Speechreading: Illusion or window into pattern recognition. Trends in Cognitive Sciences. 1999; 3:310–317. [PubMed: 10431185]

Mattys SL, Bernstein LE, Auer ET Jr. Stimulus-based lexical distinctiveness as a general word-recognition mechanism. Perception and Psychophysics. 2002; 64:667–679. [PubMed: 12132766]

McGettigan C, Faulkner A, Altarelli I, Obleser J, Baverstock H, Scott SK. Speech comprehension aided by multiple modalities: Behavioural and neural interactions. Neuropsychologia. 2012; 50:762–775. [PubMed: 22266262]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746–748. [PubMed: 1012311]

Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485:233–237. [PubMed: 22522927]

Miller JL, Grosjean F, Lomanto C. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. Phonetica. 1984; 41:215–225. [PubMed: 6535162]

Möttönen R, Schürmann M, Sams M. Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. Neuroscience Letters. 2004; 363:112–115. [PubMed: 15172096]

Möttönen R, Watkins KE. Motor representations of articulators contribute to categorical perception of speech sounds. Journal of Neuroscience. 2009; 29:9819–9825. [PubMed: 19657034]

Nath AR, Beauchamp MS. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. Journal of Neuroscience. 2011; 31:1704–1714. [PubMed: 21289179]

Oden GC, Massaro DW. Integration of featural information in speech perception. Psychological Review. 1978; 85:172–191. [PubMed: 663005]

Okada K, Venezia JH, Matchin W, Saberi K, Hickok G. An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. PLOS ONE. 2013; 8:e68959. [PubMed: 23805332]

Peelle JE. The hemispheric lateralization of speech processing depends on what “speech” is: A hierarchical perspective. Frontiers in Human Neuroscience. 2012; 6:309. [PubMed: 23162455]

Peelle JE. Methodological challenges and solutions in auditory functional magnetic resonance imaging. Frontiers in Neuroscience. 2014; 8:253. [PubMed: 25191218]

Peelle JE, Davis MH. Neural oscillations carry speech rhythm through to comprehension. Frontiers in Psychology. 2012; 3:320. [PubMed: 22973251]

Peelle JE, Gross J, Davis MH. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cerebral Cortex. 2013; 23:1378–1387. [PubMed: 22610394]

Perrodin C, Kayser C, Logothetis NK, Petkov CI. Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. Proceedings of the National Academy of Science. 2015; 112:273–278.

Reisberg, D.; McLean, J.; Goldfield, A. Easy to hear but hard to understand: A speechreading advantage with intact stimuli. In: Campbell, R.; Dodd, B., editors. Hearing by eye: The psychology of lip-reading. London: Erlbaum; 1987. p. 97-113.

Romei V, Gross J, Thut G. On the role of prestimulus alpha rhythms over occipitoparietal areas in visual input regulation: Correlation or causation? Journal of Neuroscience. 2010; 30:8692–8697. [PubMed: 20573914]

Schroeder CE, Foxe J. Multisensory contributions to low-level, ‘unisensory’ processing. Current Opinion in Neurobiology. 2005; 15:454–458. [PubMed: 16019202]

Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A. Neuronal oscillations and visual amplification of speech. Trends in Cognitive Sciences. 2008; 12:106–113. [PubMed: 18280772]

Schwartz JL, Savariaux C. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLOS Computational Biology. 2014; 10:e1003743. [PubMed: 25079216]

Sekiyama K, Kanno I, Miura S, Sugita Y. Auditory-visual speech perception examined by fMRI and PET. Neuroscience Research. 2003; 47:277–287. [PubMed: 14568109]

Seltzer B, Pandya DN. Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. Brain Research. 1978; 149:1–24. [PubMed: 418850]

Skipper JI, Nusbaum H, Small SL. Listening to talking faces: Motor cortical activation during speech perception. NeuroImage. 2005; 25:76–89. [PubMed: 15734345]

Sohoglu E, Peelle JE, Carlyon RP, Davis MH. Predictive top-down integration of prior knowledge during speech perception. Journal of Neuroscience. 2012; 32:8443–8453. [PubMed: 22723684]

Sommers MS, Tye-Murray N, Spehar B. Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. Ear and Hearing. 2005; 26:263–275. [PubMed: 15937408]

Stevenson RA, Geoghegan ML, James TW. Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. Experimental Brain Research. 2007; 179:85–95. [PubMed: 17109108]

Stevenson RA, Ghose D, Fister JK, Sarko DK, Altieri NA, Nidiffer AR, Wallace MT. Identifying and quantifying multisensory integration: A tutorial review. Brain Topography. 2014; 27:707–730. [PubMed: 24722880]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America. 1954; 26:212–215.

ten Oever S, Schroeder CE, Poeppel D, van Atteveldt N, Zion-Golumbic E. Rhythmicity and cross-modal temporal cues facilitate detection. Neuropsychologia. 2014; 63:43–50. [PubMed: 25128589]

Thut G, Nietzel A, Brandt SA, Pascual-Leone A. α-Band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. Journal of Neuroscience. 2006; 26:9494–9502. [PubMed: 16971533]

Tian X, Poeppel D. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. Frontiers in Human Neuroscience. 2012; 6:314. [PubMed: 23226121]

Tye-Murray N, Sommers MS, Spehar B. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. Ear and Hearing. 2007a; 28:656–668. [PubMed: 17804980]

Tye-Murray N, Sommers MS, Spehar B. Auditory and visual lexical neighborhoods in audiovisual speech perception. Trends in Amplification. 2007b; 11:233–241. [PubMed: 18003867]

Tye-Murray N, Spehar B, Myerson J, Sommers MS, Hale S. Crossmodal enhancement of speech detection in young and older adults: Does signal content matter? Ear and Hearing. 2011; 32:650–655. [PubMed: 21478751]

Tye-Murray N, Spehar BP, Myerson J, Hale S, Sommers MS. Reading your own lips: Common-coding theory and visual speech perception. Psychonomic Bulletin and Review. 2013; 20:115–119. [PubMed: 23132604]

Tye-Murray N, Spehar BP, Myerson J, Hale S, Sommers MS. The self-advantage in visual speech processing enhances audiovisual speech recognition in noise. Psychonomic Bulletin and Review. In press.

Van Engen KJ, Phelps JEB, Smiljanic R, Chandrasekaran B. Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. Journal of Speech, Language, and Hearing Research. 2014; 57:1908–1918.

van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Science. 2005; 102:1181–1186.

van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia. 2007; 45:598–607. [PubMed: 16530232]

Vatakis A, Ghazanfar AA, Spence C. Facilitation of multisensory integration by the “unity effect” reveals that speech is special. Journal of Vision. 2008; 8:1–11. [PubMed: 18831650]

Volgushev M, Chistiakova M, Winger W. Modification of discharge patterns of neocortical neurons by induced oscillations of the membrane potential. Neuroscience. 1998; 83:15–25. [PubMed: 9466396]

Watson CS, Nichols TL. Detectability of auditory signals presented without defined observation intervals. Journal of the Acoustical Society of America. 1976; 59:655–668. [PubMed: 1254792]

Wayne RV, Johnsrude IS. The role of visual speech information in supporting perceptual learning of degraded speech. Journal of Experimental Psychology: Applied. 2012; 18:419–435. [PubMed: 23294284]

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. Polysensory interactions along lateral temporal regions evokded by audiovisual speech. Cerebral Cortex. 2003; 13:1034–1043. [PubMed: 12967920]

Yi HG, Smiljanic R, Chandrasekaran B. The neural processing of foreign-accented speech and its relationship to listener bias. Frontiers in Human Neuroscience. 2014; 8:768. [PubMed: 25339883]

Zion Golumbic E, Cogan GB, Schroeder CE, Poeppel D. Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. Journal of Neuroscience. 2013; 33:1417–1426. [PubMed: 23345218]

Zion Golumbic E, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Schroeder CE. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron. 2013; 77:980–991. [PubMed: 23473326]

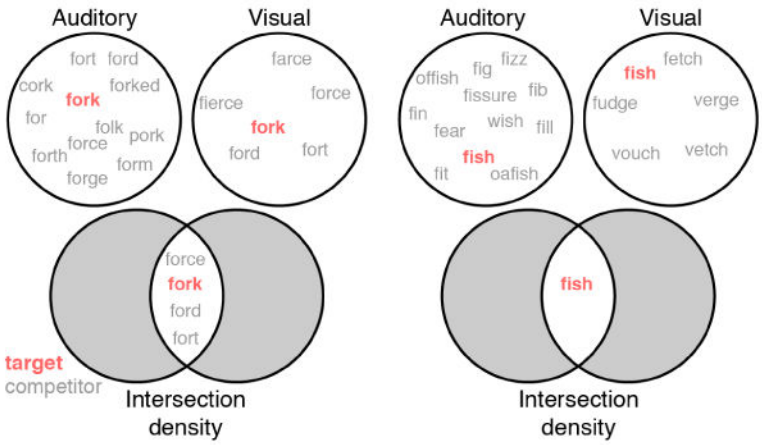

Figure 1.

Illustration of lexical neighborhoods based on auditory only, visual only, and combined audiovisual speech information (intersection density), after Tye-Murray et al. (2007b). Auditory competitors differ from a target word by a single phoneme; visual competitors differ from a target word by a single viseme.

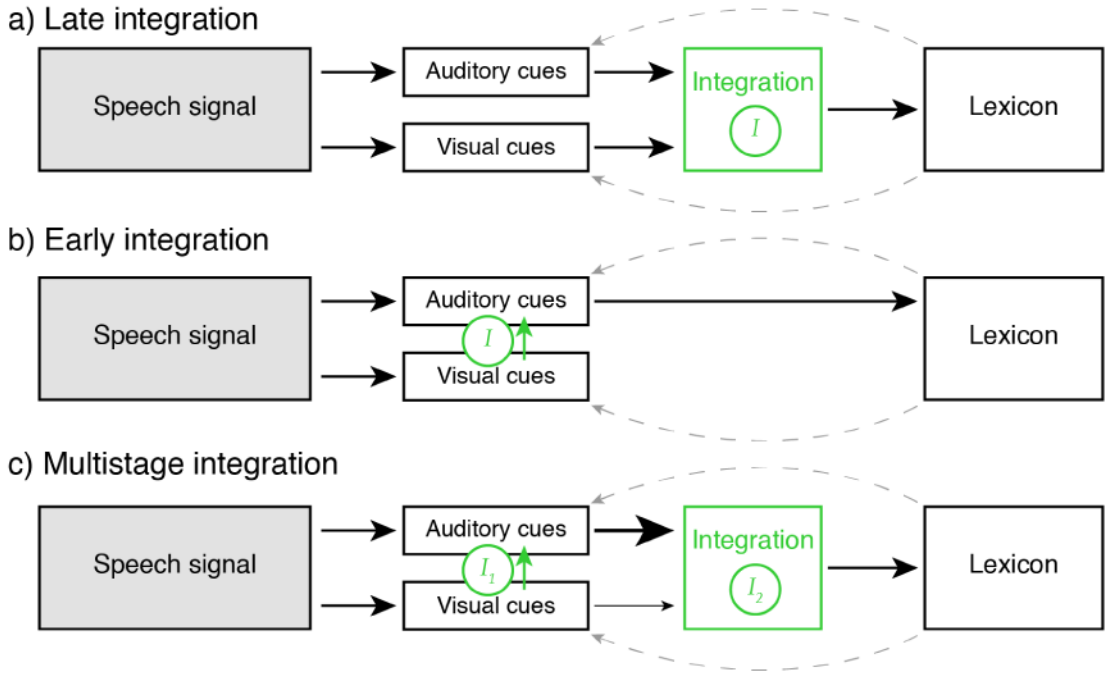

Figure 2.

Models of audiovisual speech perception. (a) A late integration view holds that multimodal integration occurs at a stage after modality-specific inputs have been processed. (b) An early integration view posits that integration happens concurrent with perception. Thus, visual information impacts the processing of auditory cues directly (there is not a pure "auditory only" representation). (c) Hybrid models allow for integration at multiple levels.

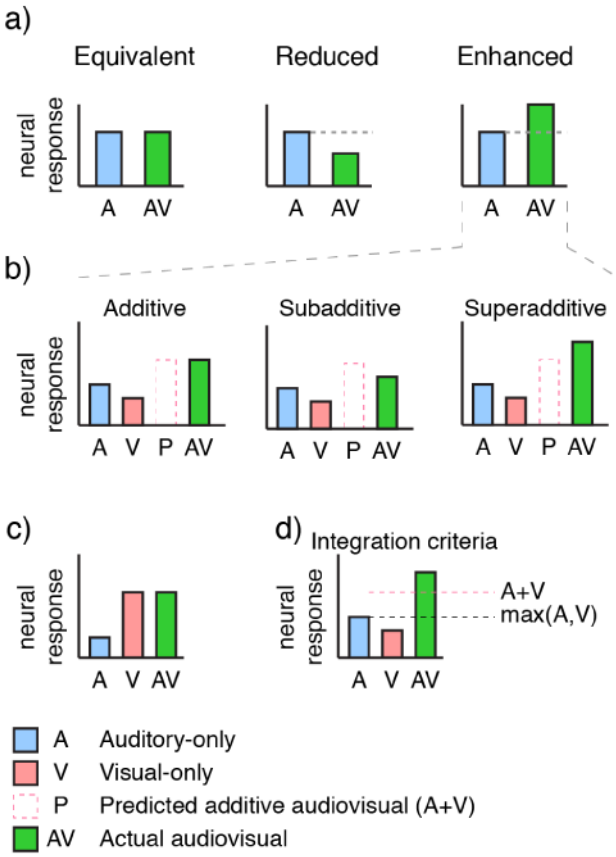

Figure 3.

Types of neural response indicating multimodal integration. (a) Responses to audiovisual speech can be categorized as equivalent to auditory-only, reduced, or enhanced. Various criteria have been used to decide whether non-equivalent responses reflect integration. (b) One frequent approach is to look at whether the response to audiovisual speech is larger than that would be expected by adding the auditory-only and visual-only responses together (characterizing enhanced audiovisual responses as additive, subadditive, or superadditive). (c) A danger of examining only auditory and audiovisual responses is that an apparent audiovisual enhancement may simply reflect a preferential response to visual stimulation. (d) For cases in which enhanced responses are observed, criteria for classifying a response as multisensory include that it be larger than the strongest unimodal response—that is, greater than max(A,V)— or that it be larger than the combined unimodal responses (larger than A+V).

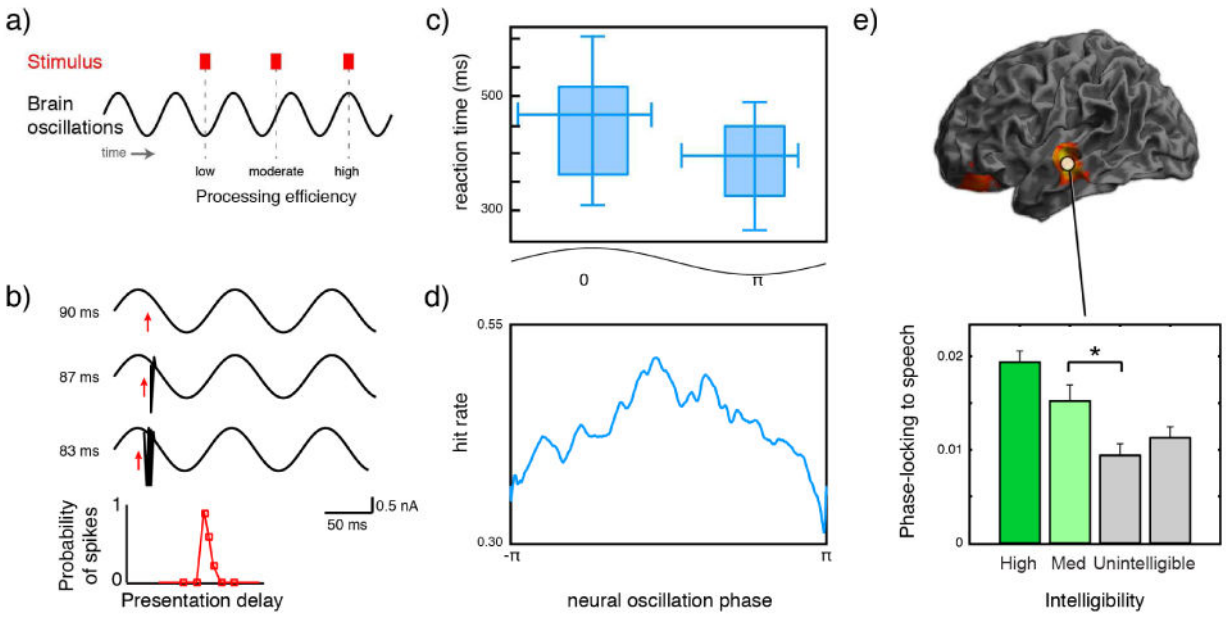

Figure 4.

Neural oscillations aid perceptual sensitivity. (a) Because oscillatory activity reflects time-varying excitability, stimuli arriving at some oscillatory phases are processed more efficiently than others. (b) Phase-based sensitivity can be examined experimentally by providing current stimulation at different phases of an oscillation (modified from Volgushev, et al., 1998). The phase of low-frequency oscillations affects behavior in numerous paradigms: (c) Reaction times (modified from Lakatos, et al., 2008). (d) Human observer accuracy in a gap detection task (modified from Henry & Obleser, 2012). (e) Low-frequency oscillations in human cortex show phase-locked responses to speech that are enhanced when speech is intelligible (modified from Peelle, et al., 2013).

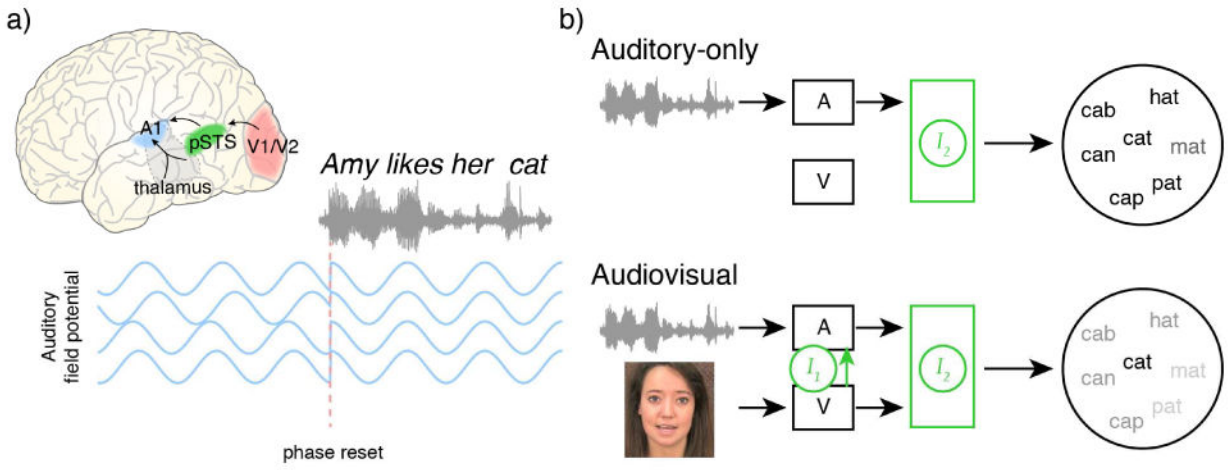

Figure 5.

Multistage integration during audiovisual speech perception. (a) Visual information (from nonspecific thalamic inputs or posterior STS) resets the phase of low-frequency oscillations in auditory cortex, increasing perceptual sensitivity. As a result acoustic cues are more salient, reducing confusability. (b) In a complementary fashion, visual speech gestures (e.g., place and manner of articulation) constrain the possible lexical candidates. In auditory-only speech (top), lexical candidates are based purely on auditory information. When visual information is available (bottom), it can act to constrain lexical identity. For example, an open mouth at the end of the word rules out the phonological neighbor “cap”, reducing the amount of lexical competition.