BRILL

Seeing and Perceiving 24 (2011) 513–539

SEEING AND PERCEIVING

brill.nl/sp

Review

Some Behavioral and Neurobiological Constraints on Theories of Audiovisual Speech Integration: A Review and Suggestions for New Directions

Nicholas Altieri [1,*], **David B. Pisoni** [2] **and James T. Townsend** [2]

[1] Department of Psychology, University of Oklahoma, 3100 Monitor Ave., Two Partners Place, Suite 280 Norman, OK 73072, USA

[2] Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405, USA

Received 10 August 2010; accepted 3 July 2011

Abstract

Summerfield (1987) proposed several accounts of audiovisual speech perception, a field of research that has burgeoned in recent years. The proposed accounts included the integration of discrete phonetic features, vectors describing the values of independent acoustical and optical parameters, the filter function of the vocal tract, and articulatory dynamics of the vocal tract. The latter two accounts assume that the representations of audiovisual speech perception are based on abstract gestures, while the former two assume that the representations consist of symbolic or featural information obtained from visual and auditory modalities. Recent converging evidence from several different disciplines reveals that the general framework of Summerfield's feature-based theories should be expanded. An updated framework building upon the feature-based theories is presented. We propose a processing model arguing that auditory and visual brain circuits provide facilitatory information when the inputs are correctly timed, and that auditory and visual speech representations do not necessarily undergo translation into a common code during information processing. Future research on multisensory processing in speech perception should investigate the connections between auditory and visual brain regions, and utilize dynamic modeling tools to further understand the timing and information processing mechanisms involved in audiovisual speech integration.

Keywords

Audio-visual speech perception, McGurk effect, multisensory enhancement

* To whom correspondence should be addressed. E-mail: nick.altieri@ou.edu

DOI:10.1163/187847611X595864

1. Introduction

The ability to obtain visual information from the movements of a talker's articulators through lip-reading is important for both normal-hearing and hearing-impaired listeners when perceiving speech (Erber, 1969; Sumby and Pollack, 1954). Sumby and Pollack (1954) demonstrated over 50 years ago that visual information in speech perception enhances accuracy scores across a wide range of signal-to-noise (S/N) ratios and that the proportion of audiovisual gain generally remains constant across different S/N ratios. Another illustration of the effect of visual information in speech perception is the well-known *McGurk effect*, a perceptual fusion that occurs when the auditory and visual signals are mismatched (McGurk and Macdonald, 1976). Specifically, the McGurk effect occurs when incongruent audiovisual information, such as an auditory /ba/ combined with a visually articulated (ga), yields a novel fusion of the two streams; in this case the fusion typically leads to the percept of /da/ (see Fowler and Dekle, 1991, for further illustrations). While many phenomena in audiovisual speech perception have been extensively researched, the neuro-cognitive mechanisms that operate on auditory and visual speech inputs during the integration process have yet to be clarified. Broadly speaking, two important issues need to be addressed.

One important question relates to the nature of the neuro-cognitive processes involved in multisensory integration in speech perception. This includes a formal/mathematical description characterizing how the time-varying dynamics of the audiovisual integration processes operate (e.g., Altieri, 2010; Altieri and Townsend, under review).

Another related question, and a significant focus of this article concerns the representations (i.e., phonetic, gestural, etc.) upon which the neuro-cognitive system operates during the integration process. Over the past several decades, several theoretical explanations have been proposed to account for such phenomena in audiovisual speech perception. In a seminal contribution to theoretical discussion on representational issues in multisensory processing, Summerfield (1987) discussed several accounts of multisensory integration in speech perception including: (1) integration of discrete phonetic features in which information about place (e.g., bilabial or velar) is obtained from the visual modality while information about manner (e.g., voiced or voiceless) is obtained from the auditory modality, (2) vectors describing the values of independent acoustical and optical parameters, (3) the filter function of the vocal tract and (4) articulatory dynamics of the vocal tract. More detailed descriptions of these accounts will be presented in the following sections.

1.1. General Background

After briefly reviewing the theoretical accounts outlined by Summerfield (1987), we propose a new theoretical framework. First, we should mention that an inherent core assumption in each of Summerfield's proposed accounts is that auditory and visual unisensory information is translated into a common code prior to the conflux of the streams (see also Rosenblum, 2005). We argue that this assumption is unnecessary

for describing multisensory phenomena. In lieu of this assumption, we propose that available evidence is most compatible with an information processing model in which temporally congruent visual information facilitates processing in auditory regions of the brain.

The framework proposed in this article is therefore consistent with Summerfield's accounts ((1) and (2)), which assume the integration of modality-specific information. We make the additional assumption that neural circuitry shares information across sensory modalities through cross-channel connections without translating the speech information into a common code. Finally, we discuss the importance of optimal timing between incoming auditory and visual speech information in the integration process.

In fact, recent modeling work carried out by Altieri and Townsend provides converging behavioral evidence for the view that auditory and visual information are processed in parallel with cross-modal interactions (Altieri, 2010; Altieri and Townsend, under review). Altieri and Townsend (under review) carried out two audiovisual speech identification tasks and computed empirical survivor functions from the reaction-time (RT) data. The RT distributions were fitted to several integration model architectures, including parallel models with separate decisions (with first-terminating *vs.* exhaustive decision rules), as well as coactive models (Townsend and Nozawa, 1995; Townsend and Wenger, 2004, for a tutorial on the Double Factorial Paradigm). Coactive models assume that auditory and visual information is combined prior to the decision stage, as opposed to parallel models, which assume separate auditory and visual decision processes. Overall, the data provided strong evidence for parallel processing with separate auditory and visual channels.

In addition to addressing the architecture question, integration efficiency was measured using the *capacity coefficient* (see Townsend and Nozawa, 1995). Computing the capacity coefficient in audiovisual identification tasks is straightforward. The RT distribution obtained from the audiovisual trials (numerator) is compared, *via* a ratio, to the sum of the RT distributions from all of the auditory-only and visual-only trials (denominator). The sum in the denominator corresponds to the predictions of an independent parallel processing model. If 'more work' is completed (i.e., faster RTs) in the audiovisual condition relative to the independent parallel model prediction, then the capacity coefficient is greater than 1, and integration is deemed efficient. Facilitatory interactions between auditory and visual channels can produce such an outcome (Eidels *et al.*, 2011). Conversely, an observation of limited capacity (ratio less than 1) suggests inefficient integration. In this latter case, inhibitory interactions between channels are usually responsible for the violation of independence. Data from an identification task using three auditory S/N ratios (quiet, −12 and −18 dB SPL) provided evidence for limited capacity, and hence inefficient integration, for high auditory S/N ratios. Interestingly, efficient integration was observed when lower S/N ratios were used (see also Ross *et al.*, 2007). These behavioral data are consistent with a parallel processing model

with cross-channel connections. This interpretation helped motivate our framework of audiovisual speech integration, which assumes separate auditory and visual processing pathways. As such, we will review recent neural data characterizing the connections between auditory and visual processing circuits involved in multimodal speech integration.

The remainder of this article will consider the implications of recent clinical and behavioral studies involving normal and hearing-impaired listeners as well as recent evidence from neuroscience to offer an updated appraisal of theoretical accounts of audiovisual integration in speech perception. As we shall see, several implications derived from recent empirical studies in audiovisual speech integration bear directly on theoretical issues regarding accounts of multi-sensory integration. Before turning to these issues directly, brief descriptions of current models of audiovisual integration will be provided below. It should be emphasized that these models are inadequate in many respects because they do not provide precise descriptions of the neural or representational basis governing multimodal perception.

A considerable amount of theoretical research on audiovisual speech perception has sought to explain how the auditory and visual cues are weighted and combined in speech recognition tasks. The Fuzzy Logic Model of Perception (FLMP) (Massaro, 1987, 2004) is an example of an approach that assumes that listeners combine auditory and visual information *via* Bayesian inference in an optimal manner during speech perception. FLMP utilizes a mathematical formula similar to Luce's well-known choice rule (Luce, 1986) to describe how auditory and visual cues are extracted and combined multiplicatively and independently and divided by the sum of the weights for the alternative stimuli (see Massaro, 2004, for details; see Note 1). An alternative modeling approach was developed by Braida (1991) to account for perceptual weighting across auditory and visual modalities. The approach is based upon concepts derived from multidimensional signal detection theory (Borg and Lingoes, 1987). Braida's Pre-Labeling Integration Model (PRE) of consonant identification assumes that auditory and visual information are accrued and assigned discrete labels such that both modalities are recognized optimally. It is generally agreed that the Bayesian inspired FLMP and the PRE model adequately fit audiovisual accuracy and confusion data (Grant, 2002; Grant and Seitz, 1998; Grant *et al.*, 2007; Massaro, 2004; Massaro and Cohen, 2000).

Nonetheless, neither the Bayesian approaches nor PRE address the issues concerning representations. Secondly, neither approach described the neuro-cognitive mechanisms involved in audiovisual speech integration. Still, these models do offer some theoretical insight into how the integration processes might function. One might, for instance, conceptualize both the Bayesian approaches and PRE as accounts of audiovisual speech integration in which the extraction of features occurs in separate auditory and visual pathways (i.e., more in line with Summerfield's accounts (1) and (2) above). We shall now return to a more complete description of the processing frameworks of audiovisual speech perception put forth by Summerfield (1987).

2. Summerfield's Accounts of Integration

The four theoretical accounts proposed by Summerfield (1987) were: (1) integration of discrete phonetic features, (2) vectors describing the values of independent acoustical and optical parameters, (3) the filter function of the vocal tract and (4) articulatory dynamics of the vocal tract. Summerfield's first account (1) refers to the integration of phonetic/phonological representations of the auditory and visual components of the speech signal. This account, known in the AV literature as Visual Place Articulatory Manner (VPAM), assumes that the cognitive system translates auditory and visual information into discrete symbolic-phonetic features prior to integration. Summerfield's second account (2) also proposes that modality-specific auditory and visual speech features are integrated. This framework further assumes that exemplars of auditory and visual spectral information (see Klatt, 1979), rather than more abstract idealized context-free phonological features, are stored in memory and matched against information obtained from incoming speech signals.

Summerfield's third (3) and fourth (4) accounts offer a contrasting perspective by assuming that sensory-motor gestural properties rather than acoustic-phonetic information is combined during perceptual analysis (Fowler and Dekle, 1991; Galantucci *et al.*, 2006; although cf. Scott *et al.*, 2009, for a recent review of gesture based theories). The third account assumes that neuro-cognitive representations of auditory and visual speech information consist of hypothetical vocal tract configurations, that are most likely to have produced the utterance. Finally, Summerfield's fourth account assumes that articulatory dynamical information obtained from the auditory and visual channels is integrated and combined together. Figure 1(a) shows a systems level depiction of the modality-specific accounts of audiovisual speech integration followed by the gestural-based theories of audiovisual speech integration (b).

We argue here that the evidence to date obtained from a wide range of studies supports modality-specific theories of integration (accounts (1) and (2)) more than gestural-based theories (accounts (3) and (4)). As such, behavioral, clinical, neural, and new research findings investigating timing aspects of audiovisual integration will be provided as evidence consistent with feature-based frameworks. First, in the following section, we will present some evidence that has been used as support for the framework that integration in speech perception operates on gestural information common to both modalities. We shall then turn to evidence for the feature-based accounts.

3. Preliminary Evidence for Gestural Representations

Summerfield (1987) proposed two accounts of audiovisual integration in which the representations operated upon by the cognitive system consist of articulatory gestures: the filter function of the vocal tract and articulatory dynamics of the vocal tract. The former account is closely related to the motor theory of speech perception (Liberman and Mattingly, 1985; see also Galantucci *et al.*, 2006), which assumes

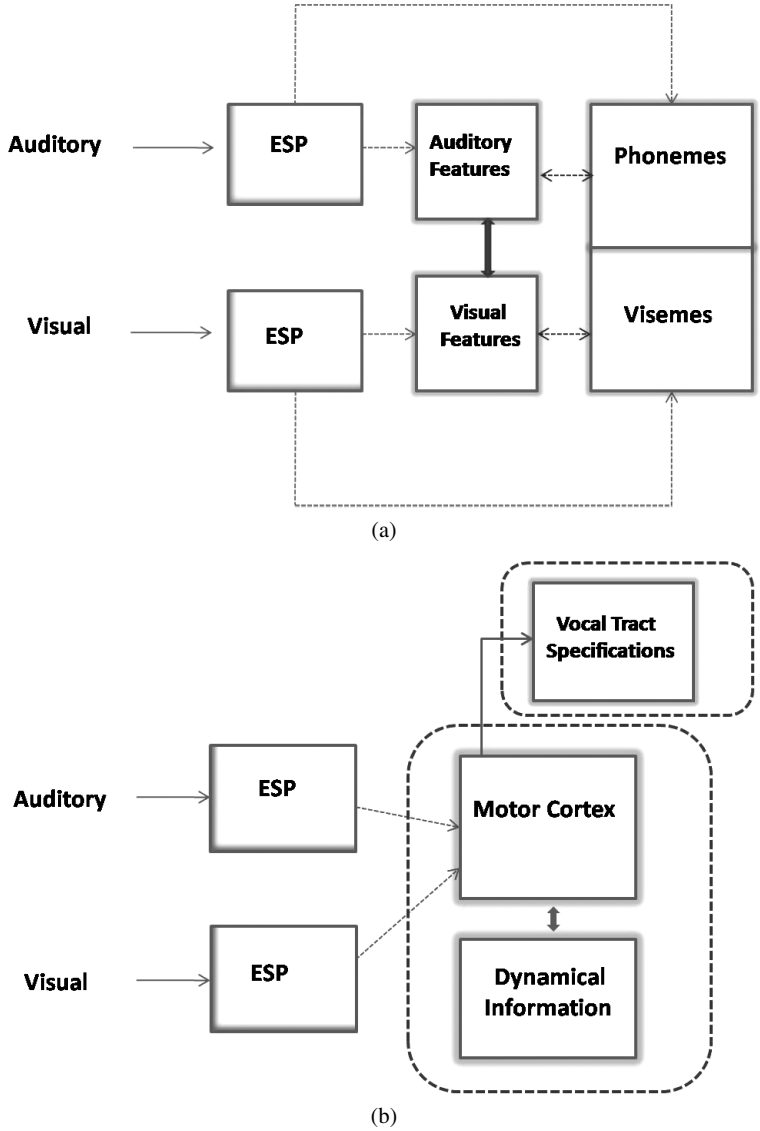

Figure 1. (a) This portion of the figure illustrates modality specific theories of integration (accounts (1) and (2)). Auditory and visual information undergo early sensory processing before translation into modality specific features relevant to spoken language. Depending on the account, modality specific information can be translated directly into phonemes or visemes, or alternatively, translated into spectral (or featural) information first before undergoing translation into higher order and more abstract units such as discrete, time invariant, phonological units. (b) Gestural theories of integration are illustrated here (accounts (3) and (4)). After early sensory encoding, relevant auditory and visual information are translated into gestural/motor codes, and subsequently mapped onto hypothetical vocal tract configurations, or alternatively, directly perceived as dynamic events. This theory does not rule out the possibility of the extraction of phonological information, although it assumes that phonological extraction would be post-perceptual or irrelevant for many aspects of real-time language processing.

that listeners generate/construct hypotheses regarding the configuration of the vocal tract responsible for producing the utterance using available auditory plus visual information obtained from lip-reading. The perceptual primitives in the articulatory dynamics framework are not parameters describing the configuration of the vocal tract *per se*, but are rather the time-varying kinematics and movements of the sound source of the utterance 'perceived directly' by listeners (see Fowler and Rosenblum, 1991, for an account of direct realism). Although these two metrics differ somewhat in their explanation of the elementary perceptual processes in audiovisual integration, they are similar by virtue of the assumption that the information operated upon by the cognitive system is independent of any particular sensory modality. Rosenblum (2005) (see also Green and Miller, 1985) argued, using gestural theories as a foundation, that behavioral and neuroimaging evidence obtained from audiovisual perception studies supports the theory that modality-independent or *amodal* information is integrated in the early stages of speech perception. The cognitive system extracts amodal gestural information from the auditory and visual channels in which this information, existing in a common articulatory code, is available for cross-modal integration prior to phonetic or word recognition.

In a study of cross-modal and talker-familiarity effects on speech recognition, Rosenblum *et al.* (2007) exposed participants to visual-only speech from a single talker for one hour and subsequently presented them with an auditory-only speech-in-noise sentence recognition task. The talker in the auditory-only recognition task was either the same talker as in the visual-only lip-reading component of the task, or a different talker. A statistically significant increase in auditory-only recognition accuracy was observed when the talker was the same in both the auditory-only and visual-only conditions. The explanation of these effects proposed by Rosenblum *et al.* (2007) was that stored amodal indexical information, such as the articulatory states specific to a talker, facilitated auditory only recognition. Similar conclusions concerning gestural and modality-independent indexical information and the transfer of information across sensory modalities have been reached based on results from cross-modal matching tasks (see Lachs and Pisoni, 2004a, b).

3.1. Evidence from Neuro-Cognitive Research

Neuro-cognitive studies of audiovisual speech perception have been used to provide further support for gesture-based theories of multisensory integration. First, functional magnetic resonance imaging (fMRI) studies using silent lip-reading tasks have revealed left hemispheric activation in cortical regions involved in auditory speech processing, including temporal association areas (e.g., Calvert *et al.*, 1997; see also Calvert and Campbell, 2003). Calvert *et al.* (1997) presented subjects with visual displays of the digits 1 through 10 and required them to rehearse the stimuli sub-vocally. Sub-vocal rehearsal of the stimuli might have led to the finding of bilateral activation in visual areas and auditory processing areas. Nonetheless, Calvert *et al.*'s (1997) results provide evidence that auditory processing regions, such as the Primary Auditory Cortex (PAC) and left STS, respond to speech sig-

nals regardless of whether transmission of the signal occurs through the auditory or visual modality. This result suggests a strong association and perhaps even a shared representation for auditory and visual speech. Interestingly, Bernstein *et al.* (2002) and Hall *et al.* (2005) also found that visual-only speech activates cortical brain regions involved in multisensory processing; however, these latter studies failed to find strong evidence for corresponding activation in auditory processing circuits.

In an imaging study investigating multisensory perception and production, Skipper *et al.* (2007) observed cortical activation across left and frontal motor regions for semantically congruent and 'McGurk' style incongruent syllables. Their study required a group of participants to (1) passively watch and listen to a speaker articulating auditory-only, visual-only, and audiovisual syllables, (2) make three-alternative forced choice responses to syllables and (3) respond by saying either 'ka', 'pa' or 'ta'. The audiovisual syllables presented to the participants could either be congruent (A_PV_P, A_KV_K or A_TV_T) or semantically incongruent (e.g., A_PV_K, where the fused percept should be perceived as 'ta'). The results from Skipper *et al.*'s (2007) experiments revealed a pattern of cortical activation in temporal frontal motor and perception areas, including the left transverse temporal gyrus and sulcus and inferior frontal sulcus. This pattern was observed when audiovisual and visual syllables were passively perceived, and when the same syllables were spoken (see also Skipper *et al.* 2005). Furthermore, when Skipper *et al.* (2007) analyzed the data from the subset of the participants that perceived 'fused' audiovisual categories ($A_PV_K \rightarrow$ 'ta'), they observed that the activation profile for the fused perception of 'ta' in frontal motor areas resembled the activation profile of congruent A_TV_T more closely than the profile for A_PV_P or A_KV_K. The authors interpreted these results as evidence for shared motor and perceptual representations that were consistent with motor theories of audiovisual speech recognition, in which perception is closely associated with a motor plan for production.

Some of the generalizations drawn from these studies should be interpreted with caution, however. While data from Skipper *et al.* (2007) and Hall *et al.* (2005) showed that audiovisual speech activates motor circuits, auditory speech generally failed to elicit activation in motor regions, and visual speech did not engage most auditory regions specific to language perception. This pattern seems to be indicative of evidence for separate representational pathways for language perception and production, especially when comparing auditory *versus* visual speech. This interpretation here appears most consistent with recent findings in the fMRI literature investigating the neural basis for language representations (Scott *et al.*, 2009).

3.1.1. Multisensory Neurons

Other researchers have interpreted data showing superadditive activation in multisensory brain regions, such as the Superior Temporal Sulcus (STS), as evidence for gestural accounts of integration (e.g., see Rosenblum, 2005). Temporally congruent auditory and visual speech stimuli sometimes elicit superadditive activation ($AV > A + V$), and subadditive activation ($AV < A + V$) or response suppression

(AV < A or V, whichever is the greatest) to incongruent audiovisual speech (Calvert and Lewis, 2004; Calvert *et al.*, 2000; Stevenson and James, 2009). The hypothesis is that multisensory neurons become activated when presented with amodal gestural information inherent in both the auditory and visual signals (Rosenblum, 2005).

Other studies, however, have questioned the use of superadditivity as a metric for multisensory integration (e.g., see Laurienti *et al.*, 2005). Likewise, it is unclear whether multisensory neurons are necessary or sufficient for audiovisual integration. Moreover, while studies involving single cell recordings in cats and non-human primates have purported to find multisensory neurons, evidence for superadditive activation in neuroimaging studies involving humans does not imply that multisensory information converges to individual neurons (see Laurienti *et al.*, 2005). In fMRI designs, the activity of individual neurons cannot be ascertained or measured directly with current techniques commonly used in cognitive neuroscience with human subjects. In fact, the observed superadditivity in a specific brain region could result from the intermingling of unisensory neurons in that region (Allman *et al.*, 2009; Bernstein, 2005; Meredith, 2002). Hence, it is possible that brain areas believed to be multisensory regions that respond uniquely or maximally to audiovisual stimuli contain large numbers of unisensory neurons that may mutually facilitate one another.

Even if superadditivity co-occurs with the presentation of audiovisual speech stimuli, it does not imply that the basic primitives of speech perception are gestures. Instead, much of the available evidence suggests the existence of highly interconnected neural pathways between auditory and visual processing areas (see Ghazanfar *et al.*, 2008). This explanation could account for Rosenblum *et al.*'s (2007) cross-modal priming results described earlier, and additionally, the data from EEG studies showing the effects of early influence from visual brain regions on multisensory speech perception (Ponton *et al.*, 2009; van Wassenhove *et al.*, 2005). Ponton *et al.* (2009), for example, used a mismatch negativity paradigm, and found evidence that the visual stimulus affects auditory processing through feedback. The authors argued that this 'modulatory feedback' most likely results from exposure to the visual component of the stimulus, rather than integration with visual linguistic features. Taken together, this evidence points to auditory-visual neural cross-channel facilitatory (Altieri, 2010; Altieri and Townsend, under review; Eidels *et al.*, 2011) interactions prior to the extraction of linguistic information rather than shared gestural representations. There is no logical reason why auditory and visual information must be translated into a common code, articulatory or phonetic, to produce cross-modal facilitation and information sharing between modalities. Visual information can facilitate auditory processing just as hearing the spoken word 'rose' evokes the visual image or even smell of a rose. While the evidence for much of the multisensory phenomena, such as multisensory regions/neurons, might be intriguing, it does not imply 'common currency' explanations of audiovisual integration, as opposed to say, simple associative learning (see Note 2).

Finally, we note here that in general, evidence from fMRI studies investigating audiovisual speech integration should to be interpreted with a certain degree of caution. One issue is that multisensory processing areas associated with audiovisual speech processing also respond to complex non-speech gestures (Puce *et al.*, 1998). The application of fMRI technology to problems in speech perception has other limitations as well, including the fact that this methodology also suffers from poor temporal resolution. The fine-grained temporal nature of speech perception makes it difficult to draw precise conclusions from fMRI designs purportedly investigating the processing dynamics of spoken language perception (see Bernstein, 2005). Rather, neuro-cognitive measures that are more sensitive to timing information, such as EEG, should be preferred (van Wassenhove *et al.*, 2007).

We will now turn to a presentation of converging evidence from behavioral and neuro-cognitive studies of audiovisual integration. The results of these investigations indicate first that extensive unisensory coding occurs prior to integration, and second, that modality-specific codes probably remain intact during integration. Using the information provided by these studies, we argue for a processing model of multisensory integration similar to the framework displayed in Fig. 1(a) together with the assumption that audiovisual facilitation/inhibition occurs though anatomical connections (from V1 to auditory processing areas) with information processing principles analogous to those found in associative or Hebbian learning (cf. Gilbert *et al.*, 2001).

4. Evidence for the Integration of Modality Specific Representations

Recent evidence obtained from studies of clinical populations provides new support for Summerfield's (1987) modality-specific processing accounts of audiovisual speech integration. A study on the McGurk effect in children with phonological disorders supports the hypothesis that modality-specific features rather than articulatory gestures are integrated and combined during multi-sensory speech processing (Dodd *et al.*, 2008). In this study, two groups of children with speech disorders, either *phonological delay* (i.e., children who show a delay in the production of age appropriate speech sounds) or *disordered phonologies* (i.e., a more profound phonological impairment than phonological delay), responded to auditory-only, visual-only and audiovisual stimuli. The auditory and visual components of the test stimuli where presented either congruently or incongruently. Participants were required to point to a picture of an item such as 'tea' or 'dough' corresponding to what they thought the talker said. In the critical experimental manipulation, the correct response on audiovisual trials sometimes required the children to perceive the McGurk fusion. For example, if visual 'key' was presented with auditory 'pea', the fusion and correct response would be 'tea'.

The accuracy scores in the auditory-only and visual-only trials did not differ as a function of subject group. Neither group of children reported many perceptual fusions in the incongruent audiovisual trials. This finding is consistent with

previous studies showing that children are less likely to report audiovisual fusions overall (see McGurk and Macdonald, 1976; and also Gori *et al.*, 2008, for a visual–haptic integration study using children). Interestingly, however, Dodd *et al.* (2008) reported an interaction in which the children with disordered phonologies were more likely to respond with the auditory component of the stimulus; visual 'key' plus auditory 'pea' evoked the response 'pea' rather than the fusion 'tea' or visual 'key'. The authors concluded that children with disordered phonologies utilized a perceptual strategy of focusing attention on auditory information on audiovisual trials.

Dodd *et al.*'s (2008) findings are consistent with the hypothesis that children with atypical phonological disorders are impaired in their ability to combine phonological information from different sources due to delays or deficits in their knowledge of phonological constraints in the auditory channel — not in their ability to extract gestural information common to each modality. Dodd *et al.* (2008) concluded that phonological rather than gestural information obtained from articulation, or knowledge of how to produce the sounds, supports lip-reading ability and multi-sensory processing in speech perception. If knowledge of how to produce the sounds were necessary for audiovisual integration in speech perception, then the disordered phonology group should not have been biased toward the auditory components of the experimental stimuli.

Studies using profoundly deaf children with cochlear implants (CIs) as participants have led to similar conclusions (Bergeson and Pisoni, 2004; Bergeson *et al.*, 2003). Bergeson *et al.* (2003) argued that evidence showing enhanced auditory-only skills in children who received CIs early and enhanced lip-reading skills in children who received them at a later age supports the hypothesis that modality-specific auditory and visual information are integrated together rather than amodal information.

Bergeson *et al.*'s explanation is supported by the finding that a CI user's ability to obtain relevant information in each sensory modality is influenced by their prior developmental history and experience with inputs from that modality. Bergeson and Pisoni (2004) also provided evidence showing that post-implantation scores in auditory-only and audiovisual conditions improve more than visual-only scores. A corollary to this argument is that if speech perception were entirely an amodal process, then improvements in accuracy scores in the auditory modality should correspond with approximately equal improvements in accuracy scores in the visual modality. It is possible, according to gestural-based theories of integration (accounts (3) and (4)), for post-perceptual processes to preferentially facilitate learning in the auditory modality without providing equivalent benefits to the visual modality. While this perspective is intriguing and cannot be ruled out entirely, it is an *ad-hoc* explanation and necessitates further evidence before being considered as a viable alternative to a modality-specific explanation.

4.1. Evidence from Neurobiology for Modality-Specific Representations

As discussed previously, obtaining neurobiological evidence to support Summerfield's (1987) proposed accounts of audiovisual speech integration has proven to be a difficult endeavor (see Dodd *et al.* 2008; Green, 1996). Neuroimaging results purporting to show evidence for anatomical location, or even the existence of multisensory neurons, are in many ways logically unrelated to representational issues. However, at least some inferences about the neuro-cognitive representations involved in audiovisual speech perception can be drawn from several recent studies.

First, one may hypothesize that if extensive processing occurs in unisensory regions associated with auditory cognition (e.g., primary auditory cortex) prior to processing in motor areas, then it would indicate that auditory features rather than amodal gestural information are extracted and combined with features obtained from the visual modality during multisensory processing. This would not by itself provide conclusive evidence for or against gesture-based representations, but it would suggest a primary role for unisensory processing during cross-modal facilitation or 'integration' of information in multisensory processing areas.

Second, one might predict segregated auditory and visual processing, even in multisensory brain regions. Indeed, recent evidence suggests that this is the case. Auditory and visual preferring neurons, in addition to bimodal neurons, appear to be highly involved in processing in multisensory regions, with auditory and visual patches acting as separate feature detectors. In fact, significant advances in knowledge of the organizational properties of multisensory brain regions have come from studies investigating single neuron or local field potentials in non-human primates (e.g., Dahl *et al.*, 2009; Ghazanfar *et al.*, 2008; Hikosaka *et al.*, 1988; see also Jones and Powell, 1970, for a seminal study).

In a study mapping out the organizational structure of neurons in the upper bank of the STS, for example, Dahl *et al.*, (2009) exposed rhesus monkeys to naturalistic auditory and visual stimuli, including primate vocalizations while obtaining recordings from bimodal and unisensory neurons. Bimodal neurons were defined as units which responded preferentially to either auditory or visual information or combined audiovisual signals and unisensory neurons were defined as units that responded preferentially or exclusively to one modality. The statistical properties of the responses revealed a topographic layout of unisensory and bimodal neurons in the superior temporal association cortex. The response profiles showed that the properties of bimodal neurons were not homogenous. While the majority exhibited additive response profiles (i.e., $AV = A + V$), a subset demonstrated superadditive response properties ($AV > A + V$), and a significantly larger subset revealed a firing pattern consistent with subadditive activation ($AV < A + V$). This heterogeneous response pattern suggests that the suppression of auditory neurons by the visual signal might even be beneficial in some circumstances, contrary to the intuition that superadditive responses are necessary for audiovisual enhancement.

Furthermore, the structure of the multisensory association area indicates that auditory neurons tend to be spatially separated from visual neurons, although 'mul-

tisensory' bimodal neurons intermingle with both. Dahl *et al.* (2009) concluded that the observed topographical structure follows the basic principles of columnar organization in other sensory cortices in which neurons are organized according to feature preference. The overall picture suggests an integration scheme in which auditory and visual neurons can separately detect unimodal features, even in multisensory brain regions. It appears that ecological auditory and visual vocalization information is shared not only through bimodal neurons, but also *via* coordination between brain regions such as STS and auditory cortex (Ghazanfar *et al.*, 2008).

4.2. Neural Oscillations

An interesting set of hypotheses also comes from analyses of the effects of neural oscillations and timing in multisensory studies. The synchronized firings of ensembles of neurons (Schroeder *et al.*, 2008) provide evidence that the PAC and associated areas are critical for audiovisual speech perception (e.g., Besle *et al.*, 2004; Pekkola *et al.*, 2005; van Wassenhove *et al.*, 2005). Schroeder *et al.* (2008) argued that multisensory facilitation occurs in the initial stages of auditory processing in the primary auditory cortex (A1). They hypothesized that visual cues obtained from lip-reading amplify auditory speech processing in A1 by shifting the phase of the oscillations such that the auditory inputs arrive when the phase is in a state of high excitability. This could have the effect of amplifying the auditory properties of the signal in cortical areas, potentially contributing to the well-known phenomenon of audiovisual enhancement (e.g., Erber, 1969; Sumby and Pollack, 1954). In contrast, if the auditory inputs were to arrive when the oscillatory phase of the neural ensembles were in a state of low excitability or at a non-resonant frequency, inhibition rather than the enhancement of the audiovisual signal should hypothetically occur.

The hypothesis regarding the role of neural oscillations and timing of auditory and visual inputs makes several unique predictions regarding the mechanisms underlying multisensory integration. One caveat, however, is that much of the current evidence for the role of neural oscillations in multisensory perception has been derived from neural recording studies using auditory and somatosensory cues in non-human primates with stimuli that are much shorter than ecologically valid audiovisual speech stimuli (e.g., Lakatos *et al.*, 2007). While behavioral support for the oscillation hypothesis is currently lacking, numerous other studies presented in the following section and results from a case study, provide additional corroborating evidence for the importance of timing in audiovisual integration.

4.2.1. Timing in Audiovisual Recognition

Although behavioral evidence in speech perception for Schroeder *et al.*'s (2008) conclusions is currently limited, their work assists in elucidating several critical issues regarding timing in the integration of audiovisual cues. Current evidence points to the significance of the temporal properties of neural responses (e.g., Ghazanfar *et al.*, 2008; Kayser *et al.*, 2010), the frequency range evoked by local field potentials during integration (e.g., Chandrasekaran and Ghazanfar, 2009), and the optimal

time window of auditory and visual events in multisensory recognition (Colonius and Diederich, 2010; van Wassenhove *et al.*, 2007).

Studies investigating cortical responses to naturally occurring vocalizations in non-human primates have observed a constrained correspondence between temporal aspects of the signal and the temporal properties of neural responses in the auditory cortex (e.g., Nagarajan *et al.*, 2002; Wang *et al.*, 1995, 2003). In a study investigating cortical responses to vocalizations in anesthetized marmosets, Nagarajan *et al.* (2002) observed a correlation between the temporal properties of the signal and neural firing, but a lack of correspondence between the spectral representation of the signal and the spectral representation of responses in A1. Other evidence revealed that A1 neurons become phase-locked to temporal components of complex signals, including vocalizations (Wang *et al.*, 1995). As such, A1 neurons were shown to be sensitive to degradation of the temporal components, but less so for spectral components of the signal. These findings could help explain why human speech recognition is strongly affected by degradation of the temporal envelope, but remains fairly robust when faced with spectral degradation (Nagarajan *et al.*, 2002; see also Shannon *et al.*, 1995).

An interesting set of studies showing the effects of distinct neural frequencies in multisensory brain regions points to a potential role for oscillatory brain activity in audiovisual integration (Chandrasekaran and Ghazanfar, 2009). These authors exposed rhesus monkeys to audiovisual video clips of rhesus vocalizations at different auditory and visual onset asynchronies, and obtained local field potentials from the STS. The authors observed that distinct frequency bands (theta: 4–8 Hz, alpha: 8–14 Hz, and gamma: $>$40 Hz) integrate face and vocalization information differently — as predicted if neural oscillations are involved in integration. The authors observed that alpha frequency bands showed enhanced power for small differences between face and voice onset time, and consistent amplification for the power of gamma frequencies across variable onset asynchronies. These data implicate the involvement of alpha oscillatory frequencies in resolving the auditory and integration window, suggesting a possible role in human speech perception.

At the behavioral level, variable onset asynchronies between the auditory and visual component of speech stimuli can disrupt integration if they are beyond a certain critical time window (Colonius and Diederich, 2010; Conrey and Pisoni, 2006; Diederich and Colonius, 2008, 2009; van Wassenhove *et al.*, 2007). Evidence for the optimal time window in the integration of auditory, visual and even tactile stimuli has been uncovered in a variety of paradigms. Experiments using information from saccades and RTs in audiovisual detection experiments (Colonius and Diederich, 2010; Diederich and Colonius, 2008, 2009), as well as EEG data showing changes in multisensory ERP components as a function of auditory-visual onset asynchronies (van Wassenhove *et al.*, 2007), have established the significance of optimal timing of the auditory and visual inputs. Colonius and Diederich (2010) speculated that the neural underpinnings of the time window of integration could be related to the extent to which visual information excites the oscillatory phase of

auditory neurons. If this were correct, then it is plausible that auditory and visual asynchronies outside of the optimal time window could disrupt oscillatory phase resetting, and consequently, audiovisual integration in speech perception (see Conrey and Pisoni, 2006).

We shall now discuss a case study of patient, who after suffering a stroke, began to perceive an auditory-visual mismatch when listening to speech and seeing a talker's face (Hamilton *et al.*, 2006). This study provides a concrete example of how a neurobiological disturbance leading to the perception of temporally mismatched auditory and visual signals could inhibit multisensory enhancement and audiovisual performance characteristically observed in speech perception.

4.2.2. A Case Study: Patient AWF

A case study of a 53 year old male (AWF), who suffered a stroke to the right and possibly left parietal lobes, revealed that audiovisual speech perception can be adversely affected if the input signals are disrupted due to a brain lesion (Hamilton *et al.*, 2006). AWF experienced an acquired deficit of audiovisual speech perception in which a temporal mismatch or asynchrony between the auditory signal and movement of the lips was perceived. AWF's perceived asynchrony led to a decrease in accuracy levels rather than enhancement in an audiovisual version of a digit span task in comparison to the auditory-only presentation condition. The patient was also slower to match words to pictures when exposed to the talker's face compared to conditions where only auditory information was available (Hamilton *et al.*, 2006).

The precise neural circuits involved in AWF's deficit are difficult to ascertain. However, it is possible, following Schroeder *et al.*'s (2008) hypothesis about the role of oscillatory phases in audiovisual integration that the visual speech inputs arrive at a sub-optimally timed point in the processing phase. As such, this could lead to audiovisual suppression rather than enhancement. It is possible then, that the neural oscillations and timing window (Colonius and Diederich, 2010; Diederich and Colonius, 2009; van Wassenhove *et al.*, 2007) in the auditory processing areas of AWF were apparently disrupted as a result of his brain lesion (see Hamilton *et al.*, 2006).

This case study of AWF provides important converging behavioral support for the hypothesis that precise timing is a necessary prerequisite for the coordinated combination of auditory and visual features across multiple neural circuits (Schroeder *et al.*, 2008; see also Chandrasekaran and Ghazanfar, 2009). As shown in AWF, even when the timing of lower-level sensory transduction in the auditory and visual channels remains intact (the bimodal components of the speech signal were presented synchronously to AWF), disrupting the timing of the arrival of visual inputs to auditory areas leads to suppression of the signal rather than enhancement. Therefore, theoretically complete accounts of audiovisual integration must describe the activation of neural mechanisms, in addition to the location involved in the timing of audiovisual inputs at all relevant processing levels.

5. Toward a Representational Framework of Audiovisual Speech Perception

The results of these recent studies have several implications for the two major frameworks of audiovisual speech perception illustrated in Fig. 1. Schroder *et al.*'s (2008) theory, experimental studies, and the case of AWF all demonstrate the importance of neural timing and synchrony in audiovisual speech perception. Cross-talk between neural circuits can become disrupted and the sharing of diverse sources of unisensory information can be seriously delayed if timing is adversely affected. The representational issue of whether the information is inherently 'gestural' or inherently 'phonemic' is de-emphasized, while matters of timing and the nature of the neural projections between brain regions become paramount.

Returning to the model theoretic discussion, converging evidence from these neuro-cognitive studies in audiovisual speech perception point to a processing model where incoming auditory and visual information undergo unisensory processing in their respective pathways but interact in a facilitatory manner at multiple processing stages if the information sources converge during an optimal time window (Schroeder *et al.*, 2008; van Wassenhove *et al.*, 2005). Consider Fig. 1(a) once again. In the framework proposed in this review, both auditory and visual information related to speech perception undergoes unisensory encoding before subsequent processing, before finally being encoded as speech or spoken language. Assuming intact neural pathways and synchronously presented audiovisual stimuli, brain regions responsible for encoding visual speech speed up and otherwise enhance auditory processing *via* connections formed by associative learning mechanisms early in development (see Wallace *et al.*, 2006).

In fact, a recent investigation demonstrated the influence of visual speech on auditory processing, even in early developmental stages. In a study with 6-month old infants, visual speech was shown to contribute to phonetic category learning (Teinonen *et al.*, 2008). Two groups of infants were exposed to speech sounds ranging on a continuum from /ba/ to /da/ paired with visually articulated tokens of (ba) or (da). In one condition (i.e., the two-category group), the visual speech corresponded to the auditory token; that is, if the auditory token was on the /ba/ side of the continuum, then the visual stimuli was (ba), and *vice versa.* In the other condition (i.e., the one-category group), only one visual phonetic category (either (ba) or (da)) was presented for the entire duration of the experiment. A stimulus-alteration preference procedure was used after exposing each group of infants to the audiovisual stimuli. Teinonen *et al.* (2008) found that only infants in the two-category condition exhibited phonetic category learning. This study provides new behavioral evidence that visual information facilitates auditory processing and phonetic category learning even in early stages of speech and language development.

While the clinical and neurophysiological evidence reviewed earlier has not entirely settled the debate regarding the nature of the neuro-cognitive representations involved in audiovisual speech perception, there have been significant advances concerning the falsifiability of Summerfield's (1987) accounts. The combined evidence suggests that extensive unisensory auditory processing occurs in relevant

speech areas, such as the primary auditory cortex, immediately prior to recognition and that unisensory visual processing occurs in separate regions, although connections between brain regions allow visual information to play a facilitatory role. Evidence further suggests that the extraction of auditory phonetic features plays a critical role in the audiovisual integration process (Dodd *et al.*, 2008), although further research is necessary before strong conclusions can be drawn on this issue.

6. Discussion

The evaluation of recent literature on audiovisual speech perception has encouraged a reappraisal of previous findings and understanding of the current state of the field, while at the same time, suggesting a revised account of the core findings. Recent applications of new tools in cognitive neuroscience, including EEG technology as well as new studies using single cell recordings in animals (see Allman *et al.*, 2009; Calvert and Lewis, 2004; Ghazanfar *et al.*, 2008; Schroeder *et al.*, 2008), have provided valuable new information about the neural underpinnings of audiovisual integration that were previously unavailable. Converging evidence across a variety of subfields made considerable contributions to assessing Summerfield's (1987) proposed accounts of audiovisual integration.

While the accounts lack precise mathematical descriptions, several novel insights have emerged that could significantly alter the way audiovisual integration is investigated in future studies. First, the framework proposed here assumes that unisensory representations derived from neural codes in auditory and visual pathways play a dominant role in multi-modal speech perception. This proposition is conceptually related to the general framework proposed in Mesulam's (1998) review of the neurobiological subsystems in human consciousness, as well as the conceptualization proposed in Bernstein's (2005) review of audiovisual integration in speech perception which assumed that sensory integration occurred in 'later' stages of perception after phonetic perception has been completed. The proposal advanced here also assumes that unisensory information relevant to speech is processed within its own unique modality-specific cortical pathway, potentially activating periphery motor circuits as well (see also Scott *et al.*, 2009), and moreover, that facilitatory connections also exist between circuits. This 'separate pathways' approach strongly diverges from alternative accounts that assume that auditory and visual information is recoded into a common gestural code or 'common currency' (e.g., Galantucci *et al.*, 2006). We discussed the argument against gestural theories that such a coding scheme (involving re-representation of the information in the respective signals after convergence onto special multisensory neurons) would be inefficient from an information theoretic standpoint (see Mesulam (1998) and Bernstein (2005) for discussion). The literature examined in this review suggests that the integrity of separate sensory inputs in both cortical and sub-cortical auditory and visual path-

ways, is an integral component in multi-modal speech perception (e.g., Dodd *et al.*, 2008; Schroeder *et al.*, 2008; Teinonen *et al.*, 2008; van Wassenhove *et al.*, 2005).

The view that auditory and visual speech information does not need to be translated or recoded into a common code during speech perception diverges significantly from several earlier accounts of audiovisual integration discussed previously (e.g., Rosenblum, 2005; Summerfield, 1987). The hypothesis that speech codes derived from the auditory and visual sensory channels are translated into a common currency, or common representational form, is a core underlying assumption in Summerfield's (1987) proposed accounts of audiovisual integration. The definition of *integration* connotes information flowing together like two rivers flowing into a common conflux; for integration to occur, it is generally assumed that the streams must somehow be coded in the same language or somehow exist in a commensurate representational format. As argued here, the common currency assumption is not necessary for multisensory enhancement or inhibition. Facilitatory cross-talk exists between different cortical and subcortical networks (e.g., Mesulam, 1998; van Wassenhove *et al.*, 2005), that likely form in early stages of neural development (cf. Wallace *et al.*, 2006). Equally as important, congruently timed information arriving from the visual sensory modality leads to facilitation from visual circuits in the cortex to auditory processing areas. This enhancement purportedly contributes to the perception of an enhanced audiovisual signal of higher clarity, resulting in greater overall accuracy and robustness in speech recognition.

Interestingly, the possibility that multisensory phenomena such as enhancement can occur without translating the unimodal sources into a common code finds support in studies of multisensory perception that do not involve speech stimuli. It is at least possible for separate neural circuits representing different sources of sensory information to facilitate or inhibit one another, perhaps through formed connections or through convergence to common processing areas, without translating the unisensory information into a common code or representational format.

An fMRI study by Österbauer *et al.* (2005) involving multisensory color-odor processing revealed that color–odor combinations that were perceived to match (i.e., red color combined with the smell of a strawberry) yielded higher levels of activation in caudal regions of the orbitofrontal cortex and insular cortex compared to color-odor combination that were perceived as a poorer match (i.e., the color green combined with the smell of a strawberry). While the authors reported that multi-sensory color–smell percepts are processed in localized brain areas, it is unnecessary to postulate that olfactory and color coding schemes must be translated into a common representation. As an aside, it does seem likely that sensory coding schemes for smell, color, sound and taste are translated into a common code in people with synaesthesia (see Marks, 1978).

Accounting for the McGurk effect (McGurk and Macdonald, 1976) within this processing framework still needs to be addressed in greater detail. One possible explanation of the perception of fused responses would be to assume that visual exemplars of (ga) become active in the visual domain. However, the projections from

the visual areas, instead of enhancing the auditory signal of /ba/, obviously provide conflicting information because visual (ga) is not commensurate on several dimensions with /ba/ in much the same way that a visual presentation of the color green is not commensurate with the smell of a strawberry. The combined information from auditory and visual circuits could then activate a neural representation perceptually closest to auditory /ba/ + visual (ga) which would often be /da/ in normal hearing adults. The link between the physical attributes of the auditory and visual speech signals, such as degree of auditory mismatch and the neural activations underlying perception of the McGurk effect, are currently being assessed. A significant set of findings addressing this issue has revealed a positive relationship between the degree of auditory-visual signal mismatch and the degree of BOLD signal activation in brain regions, particularly those implicated in audiovisual spoken language processing (Bernstein *et al.*, 2008).

While more research is necessary to deconstruct and fully understand the perceptual aspects of the McGurk effect, this explanation is perhaps most similar to the unisensory explanations proposed by Summerfield (1987) (accounts (1) and (2)), with the additional assumption that a distance similarity metric is used to map auditory dimensions onto a response category for output given visually presented stimuli (instead of a rule-based decision algorithm such as VPAM).

In summary, converging evidence from several areas of research suggests that uni-sensory theories of convergence are viable accounts of audiovisual speech integration. While this research does not immediately disprove the hypothesis that sensory codes are translated into common gestural information, several reasons were discussed for why such a coding scheme is not parsimonious. This review advances the position that facilitatory and inhibitory cross-talk between visual and auditory areas is responsible for canonical phenomena in audiovisual speech integration including audiovisual enhancement and fusion (Altieri, 2010; Altieri and Townsend, under revision; see also Eidels *et al.*, 2011). The task for future research then, is twofold: First, to continue investigating the neural-anatomical connections between auditory and visual brain regions, and second, to investigate the neural basis of the timing mechanisms involved in cross-modal integration using rigorous dynamic modeling tools.

7. Conclusions and Future Directions

Several relevant accounts of audiovisual speech perception were considered here, particularly, proposals pertaining to representations that the cognitive system operates on during cross-modal integration. These included the four accounts originally proposed by Summerfield (1987): the integration of discrete phonetic features (account (1)), vectors describing the values of independent acoustical and optical parameters (account (2)), the filter function of the vocal tract (account (3)), and finally the articulatory dynamics of the vocal tract (account (4)). After reviewing the state of the field with regard to these four proposals, current neurobiological,

clinical, and behavioral evidence was considered. The evidence at hand does not entirely settle the debate regarding the content of the auditory and visual representations of speech, although considerable progress has been made in terms of investigating claims regarding the neuro-cognitive representations and processing mechanisms involved in audiovisual speech perception.

The evidence considered here emphasizes the importance of obtaining modality-specific sensory information from the auditory and visual domains in audiovisual speech perception. It is also important to note that evidence points toward extensive processing in the primary auditory cortex in audiovisual speech perception, even prior to involvement of the STS or motor areas in the time course of spoken language processing (e.g., Schroeder *et al.*, 2008). In several respects, these recent findings support the general framework advanced by Summerfield's accounts (1) and (2); especially Summerfield's proposal that modality-specific features are integrated rather than recoded as amodal information. The information processing framework proposed here essentially builds on and extends accounts (1) and (2) (see Fig. 1(a)), although it does not necessarily distinguish between these two theories.

By Summerfield's own argumentation, account (1) was shown to be problematic because it is difficult to parsimoniously characterize the theory in terms of simple decision rules, such as integration of visual information about place information with auditory information about manner (i.e., the well-known VPAM model). Summerfield argued that for the phonetic-features hypothesis to hold, it would require considerable theoretical modifications (see Note 3). The general assumptions inherent in account (2) appear consistent with major findings in the audiovisual speech perception literature. This theoretical framework has appeal because it assumes that temporal and spectral information are stored as exemplars in working memory during processing and that fine acoustic-phonetic detail in speech is preserved and not lost or discarded. This particular assumption has gained empirical support in studies involving recognition memory of spoken words (e.g., see Palmeri *et al.*, 1993). Research over the years has, in fact, demonstrated the significance of temporal and spectral information in auditory-only speech intelligibility (e.g., see Drullman *et al.*, 1994; Drullman 1995; Shannon *et al.*, 1995).

More significantly, recent studies have revealed a tight coupling between the spectral/temporal structure of the speech signal and specific facial motion dynamics (e.g., Chandrasekaran *et al.*, 2009). These authors obtained evidence for a correspondence between temporal/spectral acoustics, and lip-motion information in an analysis of audiovisual sentence data obtained from three databases in two languages (French and English language). Their study revealed the rather striking relationship that the acoustic envelope and mouth movements are both temporally modulated in the frequency range of 2–7 Hz. The authors interpreted these findings as being highly advantageous for multisensory speech perception due to the importance of maintaining correct correspondence in audiovisual timing information (see also Kim and Davis, 2004).

Chandrasekaran *et al.* also interpreted the observation of similar temporal modulations in the auditory and visual domains as evidence for amodal theories of audiovisual integration, with temporal information being the common currency. This model-theoretic account of the data is certainly intriguing. There are nonetheless, several reasons why these findings could be accounted for more parsimoniously by modality-specific theories of integration (see Fig. 1(a)). First, while temporal modulations certainly constitute a vital component of the speech signal, auditory and visual correspondence in these domains does not exclude the importance of other modality-specific signal components. This may include modality-specific phonetic and spectral information pertinent for decoding information within the auditory component of the signal (see Shannon *et al.*, 1995). A related issue is that while redundant information may be specified by similar auditory and visual temporal modulations, it is well known that audiovisual integration mechanisms make extensive use of non-overlapping information between signals (e.g., Grant, 2002). This becomes especially true for a range of poor S/N ratios in which place and manner of articulation become degraded in the auditory domain while often remaining highly robust in the visual domain.

In spite of the above caveats, the hypothesis that cross-modal temporal modulations underlie multisensory processing, even at the neural level (Chandrasekaran *et al.*, 2009), has the potential of greatly expanding on our knowledge of the major factors contributing to speech integration. As such, temporal and spectral contribution of auditory and visual speech should continue to be investigated, especially over a wide range of auditory S/N ratios.

Despite the progress and numerous insights that have resulted from several decades of research into audiovisual speech perception, an important task for future research will be to develop improved methods for falsifying and further expanding upon the more tenable theories of audiovisual integration. The time-course of information processing in speech perception requires more thorough investigation. EEG technology constitutes an ideal tool to approach this issue because of its excellent temporal resolution (van Wassenhove *et al.*, 2005). Dynamic modeling tools have previously been applied to examine neuro-cognitive representations of speech (e.g., Grossberg *et al.*, 1997; Pitt *et al.*, 2007). While modeling paradigms such as these have been applied to auditory-only speech recognition, they should also be applicable for modeling time-based phenomena in audiovisual speech processing. Approaches such as Adaptive Resonance Theory (ART: Grossberg *et al.*, 1997), for example, may prove useful for examining the timing mechanisms involved in multimodal perception of auditory and visual speech stimuli. Parallel linear dynamic models with cross-channel interactions have provided some valuable new insights into how facilitatory and inhibitory mechanisms might be involved in multisensory integration (Altieri, 2010).

In summary, while the specific neural mechanisms involved in audiovisual enhancement still require extensive inquiry, investigations of the neuro-cognitive representations of multimodal speech have produced theoretically significant con-

clusions that can immediately lead to new directions and advances in the field with both normal hearing and hearing impaired listeners. Much work still remains to be done.

Acknowledgements

This research was supported by the National Institute of Health Grant No. DC-00111 and the National Institute of Health T32 Training Grant No. DC-00012. We would like to thank personnel in the Speech Perception Laboratory, Vanessa Taler, and two anonymous reviewers for their insightful comments and theoretical discussion on these issues.

Notes

1. In FLMP, the probability of correctly identifying stimulus /ba/, for example, instead of /da/, given the available auditory and visual information, is shown in the formula below:

$$p(/ba/|a_i, v_j) = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)}. \quad (1)$$

   The values a_i and v_j are not probabilities, but instead denote the level of auditory and visual support for a particular feature. Massaro (2004) has argued that FLMP constitutes an implementation of Bayes theorem — a theorem specifying a probability of a certain outcome (e.g., perceiving /ba/ instead of /da/) conditioned on prior knowledge or beliefs.

2. Bernstein (2005) (Bernstein *et al.*, 2004) questioned the assumptions of both motor theory (see also Scott *et al.*, 2009), along with the assumption that auditory and visual speech codes are combined 'early' into a common code. In particular, Bernstein (2005) discussed significant theoretical reasons for opposing the viewpoint that auditory and visual speech streams converge into a common neural processor in the 'early' stages of speech recognition. Convergence of distinct sensory pathways in the early stages of processing, according to some researchers (e.g., Mesulam, 1998), would be an ineffective mechanism for perceiving complex and highly variable environmental stimuli. Mesulam's argument against the convergence of distinct sources of information into a unified amodal representation is that it introduces a homunculus.
   As the argument goes, suppose that convergent neurons were necessary to represent the relevant information inherent in multimodal environmental stimuli. The homunculus would then be required to find a way to direct all important sensory information, whether it is auditory, visual or tactile, to a specific set of neurons in order for it to undergo a new representational form. This does not preclude the brain from carrying out complex tasks such as directing sensory information to multisensory neurons for re-representation, but it is problematic

since such a coding scheme would likely be far less efficient than one that represented each unimodal source of information without re-translating it once again into a common code.

3. Summerfield (1987) showed that there were several exceptions to this rule, and as a result, the theory ran into danger of becoming *ad hoc.*

References

Allman, B. J., Keniston, L. P. and Meredith, M. A. (2009). Not just for bimodal neurons anymore: the contribution of unimodal neurons to cortical multisensory processing, *Brain Topography* **21**, 157–167.

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

Auer, E. T. Jr. and Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early onset hearing impairment, *J. Speech Hearing Lang. Res.* **50**, 1157–1165.

Bernstein, L. E. (2005). Phonetic perception by the speech perceiving brain, in: *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez (Eds), pp. 79–98. Blackwell Publishing, Malden, MA, USA.

Bernstein, L. E., Auer, E. T., Moore, J. K., Ponton, C., Don, M. and Singh, M. (2002). Visual speech perception without primary auditory cortex activation, *NeuroReport* **13**, 311–315.

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

Bernstein, L. E., Lu, Z. L. and Jiang, J. (2008). Quantified acoustic–optical speech signal incongruity identifies cortical sites of audiovisual speech processing, *Brain Res.* **1242**, 172–184.

Bergeson, T. R. and Pisoni, D. B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation, in: *Handbook of Multisensory Processes*, G. A. Calvert, C. Spence and B. E. Stein (Eds), pp. 153–176. MIT Press, Cambridge, MA, USA.

Bergeson, T. R., Pisoni, D. B. and Davis, R. A. O. (2003). A longitudinal study of audiovisual speech perception by children with hearing loss who have cochlear implants, *Volta Rev.* **163**, 347–370.

Besle, J., Fort, A., Delpuech, C. and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in the human auditory cortex, *Europ. J. Neurosci.* **20**, 2225–2234.

Borg, I. and Lingoes, J. (1987). *Multidimensional Similarity Structure Analysis*. Springer, New York, NY, USA.

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

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

Calvert, G. A. and Lewis, J. W. (2004). Hemodynamic studies of audiovisual interactions, in: *Handbook of Multisensory Processes*, G. A. Calvert, C. Spence and B. E. Stein (Eds), pp. 483–502. MIT Press, Cambridge, MA, USA.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Iversen, S. D., Woodruff, P., McGuire, P. Williams, S. and David, A. S. (1997). Activation of auditory cortex during silent lip-reading, *Science* **276**, 593–596.

Calvert, G. A., Campbell, R. and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex, *Curr. Biol.* **10**, 649–657.

Chandrasekaran, C. and Ghazanfar, A. A. (2009). Different neural frequency bands integrate faces and voices differently in the superior temporal sulcus, *J. Neurophysiol.* **101**, 773–788.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech, *PLoS Computat. Biol.* **5**, e1000436.

Colonius, H. and Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis, *Frontiers Integrat. Neurosci.* **4**, 1–8.

Conrey, B. L. and Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and non-speech signals, *J. Acoust. Soc. Amer.* **119**, 4065–4073.

Dahl, C. D., Logothetis, N. K. and Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex, *J. Neurosci.*, **29**, 11924–11932.

Diederich, A. and Colonius, H. (2008). Crossmodal interaction in saccadic reaction time: separating multisensory from warning effects in the time window of integration model, *Exper. Brain Res.* **186**, 1–22.

Diederich, A. and Colonius, H. (2009). Crossmodal interaction in speeded responses: time window of integration model, *Prog. Brain Res.* **174**, 119–135.

Dodd, B., McIntosh, B., Erdener, D. and Burnham, D. (2008). Perception of the auditory-visual illusion in speech perception by children with phonological disorders, *Clin. Linguist. Phonet.* **22**, 69–82.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility, *J. Acoust. Soc. Amer.* **97**, 585–592.

Drullman, R., Festen, J. M. and Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception, *J. Acoust. Soc. Amer.* **95**, 2670–2680.

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

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli, *J. Speech Hearing Res.* **12**, 423–425.

Fowler, C. A. and Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception, *J. Exper. Psychol.: Human Percept. Perform.* **17**, 816–828.

Fowler, C. A. and Rosenblum, L. D. (1991). Perception of the phonetic gesture, in: *Modularity and the Motor Theory of Speech Perception*, I. G. Mattingly and M. Studdert-Kennedy (Eds), pp. 33–59. Lawrence Erlbaum, Hillsdale, NJ, USA.

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

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

Gilbert, C. D., Sigman, M. and Crist, R. E. (2001). The neural basis of perceptual learning, *Neuron* **31**, 681–697.

Gori, M., Del Viva, M., Sandini, G. and Burr, D. C. (2008). Young children do not integrate visual and haptic form information, *Curr. Biol.* **18**, 694–698.

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

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

Grant, K. W., Tufts, J. B. and Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing impaired individuals, *J. Acoust. Soc. Amer.* **121**, 1164–1176.

Green, K. (1996). The use of auditory and visual information in phonetic perception, in: *Speech Reading by Humans and Machines*, D. Stork and M. Hennecke (Eds), pp. 55–78. Springer-Verlag, Berlin, Germany.

Green, K. P. and Miller, J. L. (1985). On the role of visual rate information in phonetic perception, *Percept. Psychophys.* **38**, 269–276.

Grossberg, S., Boardman, I. and Cohen, M. (1997). Neural dynamics of variable-rate speech categorization, *J. Exper. Psychol.: Human Percept. Perform.* **23**, 481–503.

Hall, D. A., Fussell, C. and Summerfield, Q. (2005). Reading fluent speech from talking faces: typical brain networks and individual differences, *J. Cognit. Neurosci.* **17**, 939–953.

Hamilton, R. H., Shenton, J. T. and Coslett, H. B. (2006). An acquired deficit of audiovisual speech processing, *Brain Lang.* **98**, 66–73.

Hay-McCutcheon, M. J., Pisoni, D. B. and Kirk, K. I. (2005). Audiovisual speech perception in elderly cochlear implant recipients, *Laryngoscope* **115**, 1887–1894.

Hikosaka, K., Iwai, E., Saito, H. and Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey, *J. Neurophysiol.* **60**, 1615–1637.

Jones, E. G. and Powell, T. P. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey, *Brain* **93**,793–820.

Kaiser, A. R., Kirk, K. I., Lachs, L. and Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants, *J. Speech Lang. Hearing Res.* **46**, 390–404.

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

Kim, J. and Davis, C. (2004). Investigating the audio-visual speech detection advantage, *Speech Comm.* **44**, 19–30.

Klatt, D. H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access, *J. Phonetics* **7**, 279–312.

Lachs, L. and Pisoni, D. B. (2004a). Crossmodal source identification in speech perception, *Ecolog. Psychol.* **16**, 159–187.

Lachs, L. and Pisoni, D. B. (2004b). Cross-modal source information and spoken word recognition, *J. Exper. Psychol.: Human Percept. Perform.* **30**, 378–396.

Lakatos, P., Chen, C., O'Connell, M. N., Mills, A. and Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex, *Neuron* **53**, 279–292.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T. and Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies, *Exper. Brain Res.* **166**, 289–297.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception, *Cognition* **21**, 1–36.

Luce, R. D. (1986). *Response Times*. Oxford University Press, New York, NY, USA.

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

Marks, L. E. (1978). *The Unity of the Senses*. Academic Press, New York, USA.

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

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

Massaro, D. W. and Cohen, M. M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception, *J. Acoust. Soc. Amer.* **108**, 784–789.

Meredith, M. A. (2002). On the neuronal basis for multisensory convergence: a brief overview, *Cognit. Brain Res.* **14**, 31–40.

Mesulam, M. M. (1998). From sensation to cognition, *Brain* **121**, 1013–1052.

Nagarajan, S. S., Cheung, S. W., Bedenbaugh, P., Beitel, R. E., Schreiner, C. E. and Merzenich, M. M. (2002). Representation of spectral and temporal envelope of twitter vocalizations in common marmoset primary auditory cortex, *J. Neurophysiol.* **87**, 1723–1737.

Österbauer, R. A., Mathews, P. M., Jenkinson, M., Beckmann, C. F., Hansen, P. C. and Calvert, G. A. (2005). Color of scents: chromatic stimuli modulate odor response in the human brain, *J. Neurophysiol.* **93**, 3434–3411.

Palmeri, T. J., Goldinger, S. D. and Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words, *J. Exper. Psychol.: Learning Memory Cognit.* **19**, 309–328.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A. and Sams, M. (2005). Primary auditory cortex driven by visual speech: an fMRI study at 3T, *Neuroreport* **16**, 125–128.

Pitt, M. A., Myung, J. I. and Altieri, N. (2007). Modeling the word recognition data of Vitevitch and Luce (1998): is it ARTful?, *Psychonom. Bull. Rev.* **14**, 442–448.

Ponton, C. W., Bernstein, L. E. and Auer, E. T. (2009). Mismatch negativity with visual-only and audiovisual speech, *Brain Topography* **21**, 207–215.

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

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

Rosenblum, L. D., Miller, R. M. and Sanchez, K. (2007). Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects, *Psycholog. Sci.* **18**, 392–395.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments, *Cerebral Cortex* **17**, 1147–1153.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S. and Puce, A. (2008). Neuronal oscillations and visual amplification of speech, *Trends Cognit. Sci.* **12**, 106–113.

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

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues, *Science* **270**, 303–304.

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.

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

Sommers, M., Tye-Murray, N. and Spehar, B. (2004). Time-compressed visual speech and age: a first report, *Ear and Hearing* **25**, 565–572.

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

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

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

Teinonen, T. Aslin, R. N., Alku, P. and Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month old infants, *Cognition* **108**, 850–855.

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

Townsend, J. T. and Wenger, M. J. (2004). The serial-parallel dilemma: a case study in a linkage of theory and method, *Psychonom. Bull. Rev.* **11**, 391–418.

van Wassenhove, V., Grant, K. W. and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech, *Proc. Natl. Acad. Sci. USA* **102**, 1181–1186.

van Wassenhove, V., Grant, K. W. and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception, *Neuropsychologia* **45**, 598–607.

Wallace, M. T., Carriere, B. N., Perrault, T. J., Vaughan, J. W. and Stein, B. E. (2006). The development of cortical multisensory neurons, *J. Neurosci.* **15**, 11844–11849.

Wang, X., Lu, T. and Liang, L. (2003). Cortical processing of temporal modulations, *Speech Comm.* **4**, 107–121.

Wang, X., Merzenich, M. M., Beitel, R. and Schreiner, C. E. (1995). Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics, *J. Neurosci.* **74**, 2685–2706.