4 Phonetic Processing by the Speech Perceiving Brain

LYNNE E. BERNSTEIN

As a consequence of developments in non-invasive methods for studying brain function, the underlying neural mechanisms of speech perception are being localized spatially and temporally. Theoretical issues that until recently were addressed almost solely with behavioral evidence can now be addressed in relation to functional neuroanatomy and neurophysiology in healthy behaving humans. Some of the recent findings were not anticipated by the behaviorally-based theories. This should not be surprising, as the brain mechanisms responsible for processing speech are complex and non-linear: the expectation that behavioral evidence could adequately predict functional neuroanatomy and neurophysiology would be overly optimistic (Friston et al., 1996; Picton et al., 2000).

This chapter focuses on two central theoretical issues concerning phonetic processing. The first is whether the phonetic attributes of speech stimuli are processed by a cortical system exclusively specialized for speech. The second is whether audiovisual speech processing relies on early neuronal convergence of phonetic information. The findings discussed here support the following views: First, not all of the cortical areas that process speech are specialized for speech stimuli. Second, extensive unisensory processing precedes the binding of auditory and visual speech representations. Thus, a single phonetic processing area that is independent of sensory modality appears not to have been implemented in the speech perceiving brain.

4.1 Speech Processing along the Bottom-Up Cortical Pathways

Whether speech is processed by a specialized neural system, as opposed to a general purpose auditory system, is the subject of a longstanding debate in the speech perception literature. Liberman and Whalen (2000) reviewed the issue and framed it elegantly, opposing what they called *horizontal* versus *vertical* theories of speech perception. The former generally posit that speech stimuli are first processed by general auditory mechanisms and then are passed on to linguistic ones. The latter posit that:

> the biological roots of language run deep, penetrating even the level of speech and to the primary motor and perceptual processes that are engaged there. Seen from that perspective, speech is a constituent of a vertically organized system, specialized from top to bottom for linguistic communication. (p. 187)

Top to bottom is not a neuroanatomically precise description. In translating the phrase into neuroanatomical terms, a sensible assumption would be that *top to bottom* is relevant to cortical-level neural processing, not to the periphery (the ear and the eye), nor to the subcortical structures that intervene between the periphery and the cortex. From the ear to the auditory cortex, the speech signal is processed subcortically by the brainstem and thalamus. However, the processing at subcortical stages is most likely general to all auditory stimuli (Scott & Johnsrude, 2003). Likewise, specialization for visual speech perception (if such exists) is unlikely prior the level of the cerebral cortex.

4.1.1 The primary areas

Hearing and vision each have their own obligatory primary entry levels into cortex. The primary areas are counted as the first cortical synaptic levels. For hearing, this is the primary auditory cortex (Kaas & Hackett, 2000), also referred to as *core*, and *Brodmann area (BA) 41* (Brodmann, 1909). For vision, the primary visual cortex is V1 (Felleman & Van Essen, 1991) or BA 17 (see Figure 4.1; note that BA 41 is actually approximately in the transaxial plane but is shown externally in the figure).

Early levels of cortical processing are conventionally thought to be the first three levels of the bottom-up cortical synaptic hierarchy. In general for the auditory and visual systems, the primary unisensory cortical sensory areas project to unisensory association areas at the next and higher synaptic levels, and those areas project to other unisensory association areas at yet higher levels of the cortical synaptic hierarchy (Felleman & Van Essen, 1991; Kaas & Hackett, 2000; Mesulam, 1998). But, it should be noted, processing is not strictly serial: cortical areas become concurrently active as processing proceeds temporally (e.g., Eggermont & Ponton, 2002).

Speech information enters the auditory and visual cortical processing pathways at the same locations as nonspeech information. Primary sensory areas comprise finely tuned neurons that process elementary stimulus properties, whereas higher levels represent information with coarsely tuned neurons that are more specialized, depending on the stimulus type (Mesulam, 1998). For hearing, the primary areas process the elementary features that include pitch, temporal properties, and intensity (Eggermont & Ponton, 2002). The primary areas of the visual system process the elementary features including color, form, motion, size, and depth (Bartels & Zeki, 1998).

Studies of auditory processing using intracortical recordings in humans (Steinschneider et al., 1999) and functional brain imaging (with either functional magnetic resonance imaging (fMRI) or position emission tomography (PET)) support the generalization that speech is not preferentially processed by the primary auditory cortex (Binder et al., 2000; Celsis et al., 1999; Huckins et al., 1998; Scott &

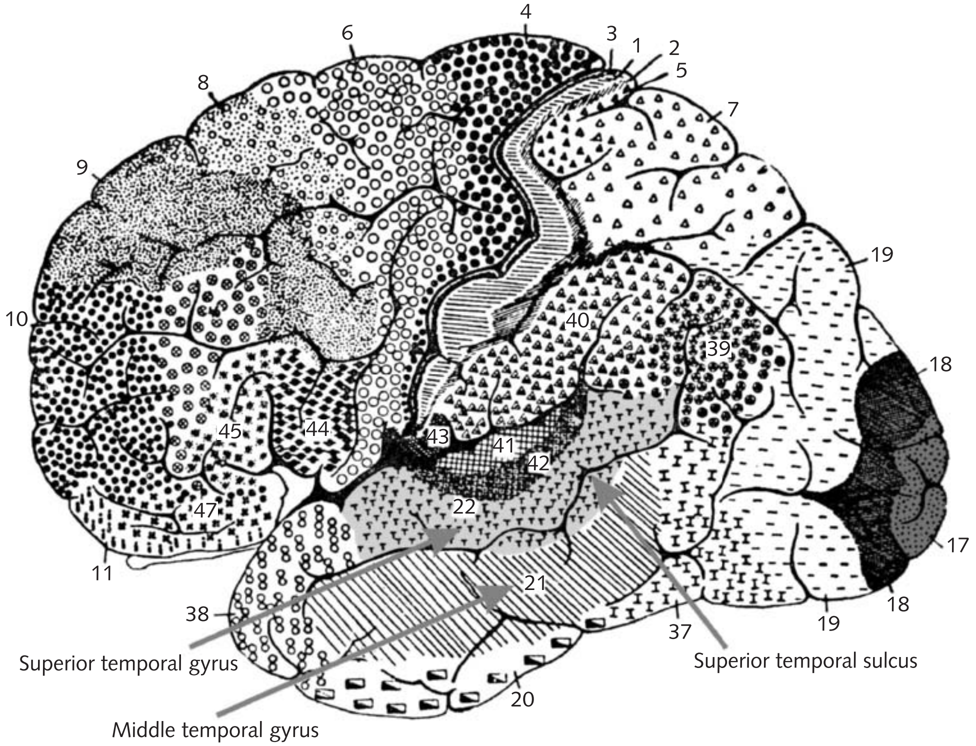

Figure 4.1 Brodmann (1909) areas of the human brain and anatomical labels.

Johnsrude, 2003; Scott et al., 2000). The primary visual area, V1, has not been studied specifically with regard to the distinction between speech and nonspeech stimuli, but there is no expectation that it would perform processing specific to speech features, just as it is not specialized for other complex visual stimuli (Mesulam, 1998).

4.1.2 The early unisensory association areas

Various animal and human studies confirm that the bottom-up flow of all auditory and visual stimulus information, including speech, is from the primary cortical areas to unisensory association areas (Bartels & Zeki, 1998; Felleman & Van Essen, 1991; Kaas & Hackett, 2000; Mesulam, 1998). The second auditory level, the first association area, designated BA 42 (see Figure 4.1), is considered to be homologous to the monkey auditory belt area, which neuronal tracer studies show receives its input from the auditory core (Hackett, Stepniewska, & Kaas, 1998). This level has been shown to be insensitive to speech versus nonspeech contrasts in humans (e.g., Belin et al., 2000; Binder et al., 2000; Eggermont & Ponton, 2002; Liégeois-Chauvel et al., 1999; Scott & Johnsrude, 2003; Wise et al., 2001).

At the third synaptic level, the primary auditory cortex and belt are surrounded by a more extensive parabelt area, designated 22 by Brodmann (1909) (Figure 4.1), and it extends onto the lateral surface of the superior temporal gyrus. Intracortical recordings in humans of evoked potentials during pre-surgical studies of epilepsy patients have shown transcortical passage of activation, which confirms that this area is hierarchically connected to the previous synaptic levels (Howard et al., 2000; Liégeois-Chauvel et al., 1994). Recent monkey studies focused on the belt and parabelt have revealed that the parabelt is also subdivided and participates in several different hierarchically organized pathways (e.g., Kaas & Hackett, 2000) that likely have different functions.

A current theory is that one auditory pathway is more concerned with category identification and the other more with location (Belin & Zattore, 2000; Kaas & Hackett, 2000; Rauschecker & Tian, 2000; Scott & Johnsrude, 2003). This possibility follows earlier work that identified *what* and *where* pathways in the visual system (Ungerleider & Haxby, 1994). Even so, although the early levels may be segregated into different paths, there is not apparently identification of specific objects at early levels. It is a general finding that the second and perhaps the third cortical synaptic levels are not specialized for any particular categories of stimulation such as speech, faces, or other objects (e.g., Binder et al., 2000; Halgren et al., 1999; Mesulam, 1998).

4.1.3 Functional evidence concerning speech processing at early areas

The results from functional brain imaging studies (fMRI and PET) are consistent with the view that the first three levels of the auditory cortex do not process phonetic stimulus attributes preferentially (Benson et al., 2001; Celsis et al., 1999; Scott et al., 2000). For example, Binder et al. (2000) presented unstructured noise, frequency-modulated (FM) tones, reversed speech, pseudowords, and words. FM tones activated the belt and parabelt areas more than did noise, but these areas were not differentially activated by speech versus FM tones.

Two types of cortical electrical data, intracortical (invasive) and scalp-recorded (non-invasive) activity, have been used to study early auditory speech processing. For example, Steinschneider et al. (1999) revealed in humans, using intracortical recordings, that voice onset time (VOT) is extremely well-represented in the primary auditory cortex and belt areas. Syllables with short VOTs produced a large electrical response, time-locked to the consonant onset, followed by a low amplitude component time-locked to voicing onset. In contrast, responses evoked by syllables with longer VOTs contained prominent components time-locked to both stimulus onset and voicing onset.

Intracortical recordings are relatively rare. Scalp-recorded event-related potentials (ERPs) obtained using electrophysiological and evoked magnetoencephalographic field (MEG) recordings afford neurophysiological measures of brain activity with high temporal resolution (< 1 ms) and, with currently evolving techniques, moderately good spatial resolution (< 10 mm). These measures are thought to mostly reflect excitatory post-synaptic potentials arising from large populations of pyramidal cells oriented in a common direction (Creutzfeldt, Watanabe, & Lux 1966; Mitzdorf, 1986; Vaughan & Arezzo, 1988).

Analyses of ERP data suggest that the origins of auditory evoked activity can be modeled successfully by dipole sources (mathematical representations of the

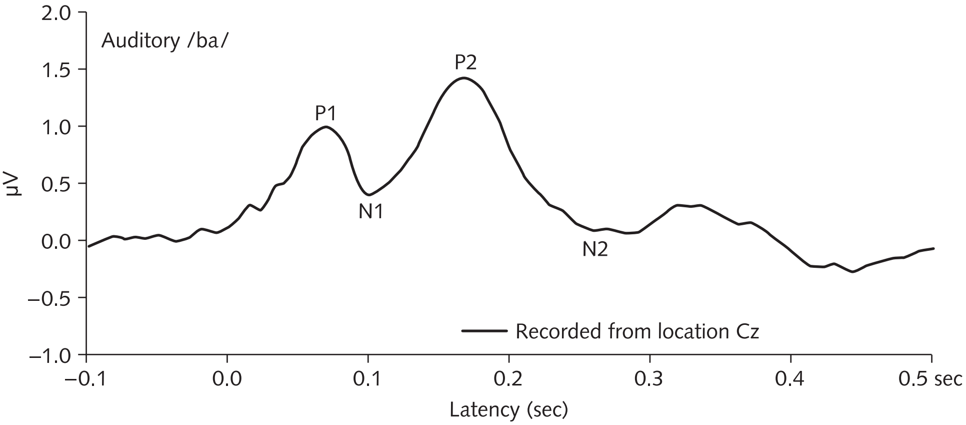

Figure 4.2 Illustration of an event-related potential to an acoustic /ba/ stimulus recorded from the scalp at the vertex location.

cortical generators) placed in the temporal lobe of each hemisphere. Positive and negative peaks in the ERP waveforms, labeled P1 (positive peak latency approximately 50 ms) and N1 (negative peak latency of approximately 100 ms) (see Figure 4.2) are represented in tangentially-oriented dipoles (i.e., sources oriented tangentially to the lateral cortical surface). That is, the P1 and N1 peaks appear to represent activity arising from the primary auditory cortex (BA 41) and the unisensory auditory association areas (belt and parabelt areas (BA 42/22)) (e.g., Knight et al., 1988; Ponton et al., 1993; Scherg & Von Cramon, 1985, 1986). These ERPs are considered to be obligatory responses to auditory system stimulation; that is, they occur in response to almost all forms of auditory stimulation, including clicks, noise or tone bursts, or speech sounds. In addition, they are obtained without requiring conscious attention to the stimuli.

While the latency and amplitude of the P1 and N1 peaks to auditory stimuli are most affected by the physical characteristics of the stimulus (e.g., duration and loudness), the later peaks (P2 – peak latency of approximately 175 ms – and N2 – peak latency about 225 ms) are more affected by factors such as arousal and attention (Näätänen & Picton, 1987), suggesting that auditory ERPs of latencies earlier than approximately 100 ms are generated at early levels of the cortical pathway. Specifically, the N1 component appears to be generated by the lemniscal pathway, which projects from the central nucleus of the inferior colliculus to the ventral division of the medial geniculate nucleus, and terminates in the primary auditory cortex, a pathway that appears to represent stimulus attributes (Ponton & Eggermont, 2001). Therefore, if there were very early specialization for speech features, it would be predicted to be reflected in the N1. Results from scalp-recorded electrophysiology studies support the prediction that N1 is not sensitive to the same contrasts observed in behavioral studies of phonetic perception.

For example, Sharma and Dorman (1999) showed with scalp-recorded ERPs that when VOT was short (/da/ stimulus), only a single distinct electrical peak component at around 100 ms (labeled N1) was obtained. For longer VOTs (/ta/ stimulus), two distinct N1 components were obtained. In addition, the discontinuity between the single component peak and the double peak coincided with the perceptual /da/-/ta/ boundary of the stimulus continuum. However, Sharma, Marsh, and Dorman (2000) in a subsequent study showed that the N1 is unlikely to be a correlate of phonetic perception. Two stimulus continua were investigated, /ba/-/pa/ and /ga/-/ka/. Perceptually, the latter results in a longer VOT boundary than the former. A single N1 peak was obtained for both continua when the VOT was < 30 ms, and a double N1 peak was obtained for both when the VOT was > 40 ms. The VOT that elicited the double N1 peak for the /ga/-/ka/ continuum occurred at 20 ms shorter VOT than required for perceptual identification of /ka/. Sharma and Dorman (2000) reported that N1 responses for a continuum from pre-voiced to voiced bilabial stops were the same for Hindi- and English-speaking listeners, even though voicing categories in the two languages do not coincide in the critical values of VOT.

Thus, the findings suggest that the responses, at around 100 ms latency and earlier, are not specialized for speech in terms of the differential coding of the phonetic properties used within language. That the scalp-recorded N1 peak has been localized independently to the auditory parabelt areas (see Eggermont & Ponton, 2002, for a review of the cortical generator sites for the early and middle auditory evoked potentials) is consistent with the conclusion that at the first two or three synaptic levels of the cortex, speech stimuli do not receive specialized neural processing, although sensitivity to speech signal characteristics is present.

4.1.4 Early visual areas

Studies of the visual pathways show that the earliest unisensory association areas are not tuned to specific visual categories (Mesulam, 1998). Specific categories such as faces and objects are not preferentially processed by early visual primary or association areas such as V1 (BA 17), V2 (BA 18), V4 (BA 19), and V5/MT (middle temporal/BA 37). For example, V5/MT is specialized for motion (Watson et al., 1993), and V4 might be specialized for color (Zeki, 2001), each at the second synaptic levels (Felleman & Van Essen, 1991). Early visual areas such as V1 and V5 are activated by visual speech stimuli (Bernstein et al., 2003; Calvert et al., 1999; Campbell et al., 2001; Ludman et al., 2000; Paulesu et al., in press), as would be expected given their role in the coding of all elementary stimulus properties. But there is not any evidence to date that these early areas are differentially sensitive to speech.

4.1.5 Phonetic processing at higher synaptic levels

The most convincing evidence for speech-specific processing has been obtained for cortical areas beyond the first three bottom-up synaptic levels. However, the evidence has not produced consensus concerning the organization of phonetic processing at the higher levels. For example, the superior temporal sulcus (which separates the superior temporal gyrus and the middle temporal gyrus, Figure 4.1), at the fourth synaptic level (Kaas & Hackett, 2000), has been shown to prefer speech to FM tones and noise (Binder et al., 2000). But the upper bank of the superior temporal sulcus has also been shown to be selective for human vocal versus non-vocal sounds, even when the vocal sounds did not contain speech (Belin et al., 2000). Spectrally inverted speech activates the left posterior superior temporal sulcus, according to Scott et al. (2000). But increasing intelligibility of speech has been shown to be associated with increasingly anterior regions along the superior temporal sulcus (Scott et al., 2000).

Binder et al. (2000) have proposed a somewhat different organization, with sensitivity to phonetic information attributed to cortex dorsal to the primary auditory cortex. However, Binder et al. (2000, Figure 9), summarizing results for the contrast between speech and nonspeech, suggest that the superior temporal sulcus is active and possibly responsible for representing temporal and spectral feature combinations. They also reported evidence that the more anterior and inferior temporal areas were more strongly activated by words than non-words, at least partly consistent with the results on anterior activity reported by Scott et al. (2000). The distinction between the posterior and anterior areas sensitive to phonetic stimuli likely results from differences in function. For example, the posterior area might be associated with the ability to repeat and represent lexical forms, whereas the anterior area might be associated with word representations and associative knowledge (cf., Hickok & Poeppel, 2000; Scott et al., 2000; Wise et al., 2001).

Unisensory visual association areas include the fusiform, inferior temporal, and middle temporal gyri. The area defined as the lateral occipital complex, located on the lateral bank of the fusiform gyrus extending ventrally and dorsally, appears to have a strong role in processing information about object shape or structure, independent of the particular visual cues to structure, and not to be differentially activated by types of visual objects (Grill-Spector, Kourtzi, & Kanwisher, 2001). At the fourth synaptic level, differential activations are observed due to complex objects versus faces (Büchel, Price, & Friston, 1998; Halgren et al., 1999; Nobre, Allison, & McCarthy, 1994; Puce et al., 1996). Face processing at this level seems to be concerned with faces as a general category, their detection and perception, but not with recognizing specific faces nor facial expressions (Tong et al., 2000). It is not known whether or not there are many category-specific areas in the visual pathway (Grill-Spector et al., 2001).

Bernstein, Auer, and Moore (2004) have speculated that visual speech is processed as a special category by the visual system. Alternatively, perhaps, this status obtains only in individuals who are proficient lipreaders (Bernstein, Demorest, & Tucker, 2000). Research on higher level vision has not resolved whether even the so-called *fusiform face area* is specialized for faces *per se* or is particularly sensitive to over-learned categories of stimulation.

Of relevance to speech, which is a dynamic visual stimulus, are studies that have investigated faces in motion. In Puce et al. (1998), a number of studies involving movements by face areas are compared. Moving eyes and mouths (nonspeech) activated a bilateral region centered in the superior temporal sulcus, 2.2–3.0 cm posterior to a region reported for lipreading in Calvert et al. (1997; see also, Bernstein et al., 2002), and 0.5–1.5 cm anterior and inferior to a region activated by hand and body movement (Bonda et al., 1996). Thus, comparison across studies seems to support specialization for phonetic versus non-phonetic visual speech processing several synaptic levels along the bottom-up visual speech pathway. But, alternatively, errors of localization during the processing and interpretation of the BOLD (blood oxygen level dependent) signal obtained during fMRI could lead to incorrect attribution of areas specialized for visual phonetic processing. Additional studies are needed that directly compare speech and nonspeech visual stimuli, controlling for a wide range of different visual stimulus properties.

4.1.6 Conclusions about phonetic processing specialization

We now return to whether "speech is a constituent of a vertically organized system specialized from top to bottom for linguistic communication" (Liberman & Whalen, 2000, p. 187). While speech is evidently part of a vertically organized system, that system does not appear to be specialized for speech at all of its levels. The pure vertical view can be saved, perhaps, by re-defining "top to bottom" beginning at a later – perhaps third or fourth – cortical level. But that maneuver does not seem consistent with the spirit of vertical theories.

On the other hand, the findings about the cortical synaptic hierarchy are also not support for the horizontalist or *auditorist* view. According to Trout (2001), "Auditorism is committed to the view that many of the distinctive achievements of speech perception . . . require *only* general auditory mechanisms, and that the auditory periphery supplies sufficient sensitivity for the analysis of the incoming speech signal" (p. 524, emphasis added).

The neural findings suggest that speech perception *does* require general purpose mechanisms, both auditory and visual, and that the periphery (ear and eye), as well as subcortical structures, *must* supply sufficient sensitivity for the information in the speech signal to be preserved for processing at higher cortical levels. But evidence for early elementary processing, and even processing similarities across species (cf., Eggermont, 2001), does not constitute evidence for the sufficiency of general purpose auditory mechanisms for phonetic perception. The findings suggest that there are higher level cortical areas that are sensitive to phonetic stimulus forms. Thus, what might be called *pure* horizontal or vertical theories do not seem to have been implemented in the human cerebral cortex. Earlier processing appears to be responsible for elementary auditory attributes and later processing appears to be more sensitive to phonetic information. Furthermore, the auditory and visual pathways appear to become more specialized at approximately the same rate from one synaptic level to the next in the bottom-up cortical pathway.

Alternatively, Bartels and Zeki (1998) have suggested that a specialized system can be defined as the direct and indirect pathways (including feedback) to the specialized areas along with those specialized areas. For example, the specialized system for processing visual motion could be regarded as the earlier visual areas V1 followed by V2 that are not specialized for motion. Furthermore, they propose that nodes or areas within a system can independently contribute to conscious perception, as, for example, when visual form is perceived from motion, but the motion of the forms is itself also perceived. Similarly, in the case of auditory speech perception, early auditory areas are shared among specialized systems such as phonetic and voice processing systems (Belin et al., 2000), allowing the listener to perceive both the message and the qualities of the talker's voice. Specialized systems thus can, and must, share less specialized cortical areas. This is a direct consequence of the hierarchical and parallel architecture of the bottom-up cortical pathways (Mesulam, 1998; Rauschecker, 1998; Zeki, 1998).

4.2 Audiovisual Speech Processing

The second main issue here is whether audiovisual (AV) phonetic processing relies on early convergence. Early phonetic convergence is taken here to imply involvement of multisensory cells at early levels of the bottom-up synaptic hierarchy that are specialized for processing both auditory and visual phonetic information. A quintessential example of neural convergence studied by Stein and colleagues (Meredith, 2002; Stein, 1998; Stein & Meredith, 1993) is the multisensory neurons in the (anaesthetized) cat superior colliculus, a subcortical structure concerned with detection and localization of events in extra-personal space. These neurons respond to more than one of auditory, visual, and somatosensory stimulation, and do so – under certain stimulus conditions – more vigorously than would be predicted by summing their unisensory responses. Meredith (2002) points out, however, that for the mammalian brain, "relatively little is known about the nature of multisensory convergence onto individual neurons and the functional architecture underlying multisensory convergence" (p. 33).[1] Convergence could alternatively involve neural networks that represent stimulus information, independent of the sensory input system. In either case, beyond the convergence process, the stimulus would be represented amodally.

4.2.1 AV speech perception

Early AV phonetic convergence has intuitive appeal, because the phonetic effects of AV processing are rarely consciously noted during everyday communication. Speech researchers themselves paid hardly any attention to AV speech perception (cf., Sumby & Pollack, 1954), until McGurk and MacDonald (1976) published their study in which mismatched auditory and visual syllables were presented. An example of the so-called *McGurk effect* is when an auditory /ba/ is dubbed to a visual /ga/, and listeners report hearing /da/. Numerous studies have replicated the McGurk effect (e.g., Green & Kuhl, 1989; Green & Norrix, 2001; Massaro, 1987; Massaro, Cohen, & Smeele, 1996; Munhall & Tohkura, 1998; Munhall et al., 1996; Saldaña & Rosenblum, 1994; Sekiyama, 1997; Walker, Bruce, & O'Malley, 1995).

The typical description or explanation of McGurk effects, expressed at the level of sub-segmental phonetic features, is consistent with a theoretical early convergence mechanism (Fowler, 2004; Green, 1998; Massaro, 1989, 1999; Schwartz, Robert-Ribes, & Escudier, 1998; Summerfield, 1987; cf., Braida, 1991). McGurk perceptual effects appear to emerge from a process that eliminates the original sensory stamp from the phonetic information, producing a transformed auditory impression. That the neural processing mechanism results in an amodal representation has seemed "uncontroversial" to some theorists (Schwartz et al., 1998).

What has been debated is the form that the amodal representation might take. Rosenblum (2002), for example, states that, "the informational metric taken at the point of speech integration is best construed as an articulation based, modality-independent form" (p. 1461) Massaro (1987) has proposed an integration process that involves independent analysis of modality specific sub-segmental features that are evaluated against abstract phoneme representations. That is, segmental representations are the abstract (hence modality-independent) products of combining features. The possibility that auditory and visual representations might bind after modality specific processing of larger patterns – beyond the level of phonetic features – has been explicitly rejected by some theorists (Summerfield, 1987; Braida, 1991).

Indeed, several types of additional behavioral evidence are consistent with early AV phonetic integration, including the following: (1) Gender incongruency between auditory and visual stimuli does not abolish the McGurk effect (Green et al., 1991); (2) Selective attention to one modality or the other does not abolish it (Massaro, 1987); (3) Explicit knowledge about incongruity between auditory and visual stimuli does not abolish it (Summerfield & McGrath, 1984); and (4) Phonetic goodness judgments can be affected by visual speech (Brancazio, Miller, & Pare, 2000). All of these effects imply that AV processing is not penetrated by high level cognition and is, therefore, an early process.

However, behavioral evidence that does not seem consistent with early AV processing also exists: (1) Large stimulus onset asynchronies between auditory and visual syllables do not abolish the McGurk effect (± 0.267 ms, Massaro et al., 1996; 180 ms, Munhall et al., 1996); (2) Reductions in the strength of the McGurk effect occur for familiar versus unfamiliar talkers (Walker et al., 1995); (3) McGurk effect strength varies across language or culture (Sekiyama & Tohkura, 1993); (4) Reductions in McGurk effect strength can be obtained as the result of training (Massaro, 1987); and (5) Visual stimuli do not selectively adapt auditory speech continua (Saldaña & Rosenblum, 1994).

AV effects involving long stimulus onset asynchronies suggest that processing latencies need not be early. Effects due to talker familiarity and culture or language suggest a role for high level cognition. Training effects could be due to changes in perception and/or attention, as well as higher-level, post-perceptual strategies. The demonstration that a visual phonetic stimulus does not selectively adapt an auditory phonetic continuum has been interpreted as evidence that auditory and visual phonetic processes do not interact early (Bernstein et al., 2004).

4.2.2 Evidence for early convergence from fMRI

As reviewed above, processing that is specifically phonetic seems to be initiated no earlier than the fourth bottom-up synaptic level of the cerebral cortex (Benson et al., 2001; Binder et al., 2000; Celsis et al., 1999; Scott & Johnsrude, 2003; Scott et al., 2000). If AV phonetic processing relies on convergence of phonetic representations, that should occur no earlier than unisensory phonetic processing.

Calvert, Campbell, and Brammer (2000) obtained response patterns to AV versus auditory-only and visual-only speech using fMRI. Congruent AV speech resulted in superadditive activation levels in the posterior ventral bank of the superior temporal sulcus relative to the sum of activation in response to auditory-alone and visual-alone speech. Incongruent AV speech produced responses lower in activation than the sum of the responses to the unisensory stimuli. This pattern was interpreted as indicative of convergence, possibly of the type observed for multisensory neurons in the superior colliculus (Stein & Meredith, 1993). Calvert et al. concluded that "these data clearly support the hypothesis that crossmodal binding of sensory inputs in man can be achieved by convergence onto multisensory cells localised [*sic*] in heteromodal cortex" (p. 655).

That conclusion is not inevitable. The superior temporal sulcus is an extremely complex and large multimodal area. It responds not only to speech but also to nonspeech motions of mouths and eyes (Puce et al., 1998). It is activated by spoken and written words (Binder et al., 2000; Fiez et al., 1996). It is activated in deaf adults viewing fingerspelling (Auer, Bernstein, & Singh, 2001). Raij, Uutela, and Hari (2000) showed that the left posterior superior temporal sulcus was activated in response to combinations of spoken and written letters of the Finnish alphabet.

In addition, the BOLD response is an indirect measure of neural activity. As a result, fMRI spatial resolution is not fine enough to obtain data on individual neurons, as is done for recordings made in animal models (Meredith, 2002). Thus, the effects reported by Calvert et al. (2000) could be due to co-mingled unisensory neurons (Meredith, 2002; Zeki, 2001). Also, fMRI temporal resolution is on the order of seconds, yet convergence based on early bottom-up phonetic processing would be predicted to occur within approximately 150 ms, in order to be consistent with the dynamics of bottom-up stimulus processing through the first several synaptic levels (Foxe & Simpson, 2002; Krolak-Salmon et al., 2001; Steinschneider et al., 1999; Yvert et al., 2001). Thus, it is not possible to know with fMRI the detailed temporal dynamics of AV processing. Another consideration is that fMRI at the strengths used with humans does not resolve the activity at different cortical layers. The consequence here is that activity in a particular area cannot be unambiguously attributed to either feedforward or feedback units.

4.2.3 Neuroanatomical problems with convergence

Convergence is also problematic given longstanding results from neuroanatomy. Classical monkey studies failed to show direct early connections between the auditory core and V1 (Jones & Powell, 1970). Mesulam (1998) summarizing the literature states that "One of the most important principles in the organization of the primate cerebral cortex is the absence of interconnections linking unimodal areas that serve different sensory functions" (p. 1023). Furthermore, this seems to be true also at the level of early unisensory association areas. According to Mesulam:

> This is particularly interesting since many of these unimodal association areas receive monosynaptic feedback from heteromodal cortices which are responsive to auditory and visual stimuli. The sensory-petal (or feedback) projections from heteromodal cortices therefore appear to display a highly selective arrangement that actively protects the fidelity of sensory tuning during the first four synaptic levels of sensory-fugal [feedforward] processing. (p. 1023)

Even at the second through fourth synaptic levels, processing should be protected from contamination by phonetic information of another input system. If this principle holds, and if phonetic processing is initiated at the fourth synaptic level, AV interaction follows unisensory phonetic processing.

However, very recent results of scalp-recorded electrophysiological studies in humans have implied that there could be auditory inputs to early visual areas (Giard & Peronnet, 1999; Molholm et al., 2002). Also, Falchier et al. (2002) showed in monkeys, using tracers injected into V1, that there was a small proportion of connections from auditory cortex to the visual periphery area of V1. There were virtually no connections to the central visual cortex. However, these auditory projections were suggested to play a role in spatial localization or event detection. Thus, this animal model does not directly support the existence of early AV phonetic convergence in humans. At present, knowledge about cortical architecture remains incomplete, and connections between early areas of sensory cortices might be found that have some functional role in AV phonetic processing. Strong evidence should be required, however, to overturn the classical conclusions about the protection against contamination of information by an alternate sensory system as articulated by Mesulam (1994, 1998). To overturn the classical view for speech requires showing that any early transcortical connections actually involve phonetic processing and not merely detection, localization, or response modulation.

4.2.4 Theoretical arguments against convergence

Outside of speech perception, convergence has been considered deficient as a potential mechanism for perceiving complex, non-invariant stimuli. Mesulam (1994, 1998) has pointed out that, in principle: (1) If a convergent cortical area were needed to represent all of the information relevant to a complex percept, then the brain (or an omniscient homunculus) would have to solve the problem of directing all of the needed information to that location for re-representation; and (2) Convergence of the type in (1) would lead to contamination of the original perceptual information.[2]

Mesulam (1998) suggests, as an alternative, areas that act as binding sites for sensory-specific representations, perhaps, by creating look-up tables or links. Mesulam employs the term *convergence* in the sense of multiple unisensory pathways feeding into the same area. But he specifically rejects the notion that complex, non-invariant information converges onto the same representation from two different sensory systems. Even at higher synaptic levels, he questions the possibility of convergence onto a common format.

Singer (1998) points out that while convergence on particular sets of neurons in a feedforward architecture is useful for rapid processing of frequently encountered stereotyped combinations of stimulus attributes, convergence is very costly in terms of the number of neurons needed and is not well-suited to dealing with varying and diverse stimulus properties. Singer proposes that the brain solves the binding problem by creating functionally coherent, dynamically created assemblies that as a whole represent particular stimulus content. These dynamic units are thought to be brought about through widespread neuronal synchronization.

Zeki (1998) points out that anatomical studies of the visual system show that there are no cortical areas to which visual pathways uniquely project (see Felleman & Van Essen, 1991), and which act as integrators of all the different visual sources (Bartels & Zeki, 1998). Moutoussis and Zeki (1997) have commented on the charm of convergence:

> To all of us, intuitively much of the most appealing solution [to the binding problem] was an anatomical convergence, a strategy by which the results of operations performed in all the specialized visual areas would be relayed to one area or a set of areas – which would then act as the master integrator and, perhaps, even perceptive areas. Apart from the logical difficulty of who would then perceive the image provided by the master area(s) . . . there is a more practical difficulty – the failure to observe an area or a set of areas that receive outputs from all the antecedent visual areas. Thus the convergent anatomical strategy is not the brain's chosen method of bringing this integration about. (pp. 1412–13)

4.2.5 *Convergence to overcome the diverse qualia of auditory and visual speech*

Nevertheless, convergence for AV speech stimuli might seem justified because of the diverse qualia of auditory and visual stimuli that seem to argue for a transformation into a common amodal format. However, researchers have noted that the biomechanical speech articulation processes that produce acoustic signals also produce optical signals. This commonality of origin has justified studies of the relationship between acoustic and optical speech signals (Jiang et al., 2002; Yehia, Kuratate, & Vatikiotis-Bateson, 1999; Yehia, Rubin, & Vatikiotis-Bateson, 1998). These studies have shown that there are consistent relationships between optical and acoustic phonetic measures.

Bernstein et al. (2004) proposed that binding of auditory and visual speech information could be accomplished by cortical networks that have learned the predictable correspondence between auditory and visual information, without re-representing the information in an amodal format. They demonstrated, using methods from multilinear regression (Jiang et al., 2002) applied to acoustic and optical speech signals, that correspondence between acoustic and optical speech signals could be established without conversion to a common metric (see also Yehia et al., 1998).

For example, Jiang et al. (2002) showed that a linear relationship could be computed between acoustic features and 3-dimensional optical measures. This demonstration was performed on a large number of nonsense syllables and several sentences spoken by four talkers. Good predictions were obtained from acoustic features to 3-dimensional optical data and vice versa. The systematic correspondence between optical and acoustic phonetic stimulus patterns could be learned by a speech perceiving brain, which would not be required to re-represent acoustic and optical patterns in terms of an amodal format. The binding problem for AV speech could be solved, perhaps, by synchrony among sensory-specific distributed representations.

4.3 General Conclusions

Phonetic perception has now taken its place as a central topic in cognitive neuroscience. Admittedly, these are early times in studying the speech perceiving brain, but the findings that have begun to emerge clearly challenge previously held views on the two issues presented here. The first main issue concerned the theoretical dichotomy between horizontal and vertical theories of speech perception (Liberman & Whalen, 2000), which motivated much debate; but the speech perceiving brain appears to be organized along neither dimension exclusively. The second main issue concerned AV speech processing research, which was seen to belong within the general area of research on neural binding mechanisms, a fact that is well-recognized within cognitive neuroscience (Calvert, 2001; Molholm et al., 2002; Mottonen et al., 2002; Raij et al., 2000). Although little is known about the neural mechanisms of AV phonetic binding, the behaviorally based accounts of early phonetic convergence seem unlikely on anatomical grounds; and theoretical considerations argue generally against neuronal convergence for binding complex non-invariant representations.

ACKNOWLEDGMENTS

The author expresses thanks to Edward T. Auer, Jr. for his comments on the chapter and his assistance with Figure 4.1. This chapter was written with support of the National Science Foundation. The views expressed here are those of the author and do not necessarily represent those of the National Science Foundation.

NOTES

[1] Bernstein et al. (2004) presented a review of the literature on the neural connections involving the superior colliculus in the monkey. Their review concluded that superior colliculus AV convergence is unlikely to be the support for the binding of AV phonetic information in humans. That is, subcortical AV convergence seems an unlikely mechanism for AV phonetic convergence at the cortical level.

[2] Of course, results from McGurk experiments seem to show contamination. But it should be noted that Sekiyama and Tohkura (1991) have shown that the McGurk effect is significantly weaker in Japanese relative to American perceivers. At the same time, critically, the Japanese results indicate separable sensitivity to visual speech information. They state

> When the stimuli . . . were composed of conflicting auditory and visual syllables, the Japanese subjects often reported incompatibility between what they heard and what they saw, instead of showing the McGurk effect . . . This implies that the visual information is processed to the extent that the audiovisual discrepancy is detected most of the time. It suggests that, for

> clear speech, the Japanese use a type of processing in which visual information is not integrated with the auditory information even when they extract some lip-read information from the face of the talker. (p. 76)

These findings suggest that sensory-specific representations are maintained by Japanese perceivers, and that AV integration of the McGurk type is therefore not obligatory nor necessarily early.

REFERENCES

Auer, E. T., Jr., Bernstein, L. E., & Singh, M. (2001). Comparing cortical activity during the perception of two forms of biological motion for language communication. In D. W. Massaro, J. Light, & K. Geraci (eds.), *Proceedings of Audio Visual Speech Perception 2001* (pp. 40–4).

Bartels, A. & Zeki, S. (1998). The theory of multistage integration in the visual brain. *Proceedings of the Royal Society of London*, Series B: Biological Sciences, 265, 2327–32.

Belin, P. & Zattore, R. J. (2000). "What," "where" and "how" in auditory cortex. *Nature Neuroscience*, 3, 965–6.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex, *Nature*, 403, 309–12.

Benson, R. R., Whalen, D. H., Richardson, M., Swainson, B., Clark, V. P., Lai, S., & Liberman, A. M. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain and Language*, 78, 364–96.

Bernstein, L. E., Auer, E. T., Jr., Moore, J. K., Ponton, C. W., Don, M., & Singh, M. (2002). Visual speech perception without primary auditory cortex activation. *NeuroReport*, 13, 311–15.

Bernstein, L. E., Auer, E. T., Jr., Zhou, Y., & Singh, M. (2003). Cortical specialization for visual speech versus non-speech face movements in color video and point lights. Cognitive Neuroscience Society, March 30–April 1.

Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual speech binding: convergence or association? In G. A. Calvert, C. Spence, & B. E. Stein (eds.), *Handbook of Multisensory Processing* (pp. 203–23). Cambridge, MA: MIT Press.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, 62, 233–52.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, 10, 512–28.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience*, 16, 3737–44.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 43A, 647–77.

Brancazio, L., Miller, J. L., & Pare, M. A. (2000). Visual influences on internal structure of phonetic categories. *Journal of the Acoustical Society of America*, 108, 2481.

Brodmann, K. (1909). *Vergleichende Lokalisatinslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: J. A. Barth.

Büchel, C., Price, C., & Friston, K. (1998). A multimodal language region in the ventral visual pathway. *Nature*, 394, 274–77.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral Cortex*, 11, 1110–23.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *NeuroReport*, 10, 2619–23.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., Woodruff, P. W., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276, 593–6.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, 10, 649–57.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M. J., & David, A. S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, 12, 233–43.

Celsis, P., Boulanouar, K., Doyon, B., Ranjeva, J. P., Berry, I., Nespoulous, J. L., & Chollet, F. (1999). Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *NeuroImage*, 9, 135–44.

Creutzfeldt, O. D., Watanabe, S., & Lux, H. D. (1966). Relations between EEG phenomena and potentials of single cortical cells. I: Evoked responses after thalamic and epicortical stimulation. *Electroencephalography and Clinical Neurophysiology*, 20, 1–18.

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, 157, 1–42.

Eggermont, J. J. & Ponton, C. W. (2002). The neurophysiology of auditory perception: from single units to evoked potentials. *Audiology & Neuro-otology*, 7, 71–99.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, 22, 5749–59.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1, 1–47.

Fiez, J. A., Raichle, M. E., Balota, D. A., Tallal, P., & Petersen, S. E. (1996). PET activation of posterior temporal regions during auditory word presentation and verb generation. *Cerebral Cortex*, 6, 1–10.

Fowler, C. (2004). Speech as a supramodal or amodal phenomenon. In G. Calvert, C. Spence, & B. E. Stein (eds.), *Handbook of Multisensory Processes* (pp. 189–201). Cambridge, MA: MIT Press.

Foxe, J. J. & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans: A framework for defining "early" visual processing. *Experimental Brain Research*, 142, 139–50.

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S., & Dolan, R. J. (1996). The trouble with cognitive subtraction. *NeuroImage*, 4, 97–104.

Giard, M. H. & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, 11, 473–90.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech* (pp. 3–25). Hove, UK: Psychology Press.

Green, K. P. & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, 45, 34–42.

Green, K. P. & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. *Journal of Experimental Psychology: Human Perception and Performance*, 27, 166–77.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, 50, 524–36.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, 41, 1409–22.

Hackett, T. A., Stepniewska, I., & Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comparative Neurology*, 394, 475–95.

Halgren, E., Dale, A. M., Sereno, M. I., Tootell, R. B., Marinkovic, K., & Rosen, B. R. (1999). Location of human face-selective cortex with respect to retinotopic areas. *Human Brain Mapping*, 7, 29–37.

Hickok, G. & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, 4, 131–8.

Howard, M. A., Volkov, I. O., Mirsky, R., Garell, P. C., Noh, M. D., Granner, M., Damasio, H., Steinschneider, M., Reale, R. A., Hind, J. E., & Brugge, J. F. (2000). Auditory cortex on the human posterior superior temporal gyrus. *Journal of Comparative Neurology*, 416, 79–92.

Huckins, S. C., Turner, C. W., Doherty, K. A., Fonte, M. M., & Szeverenyi, N. M. (1998). Functional magnetic resonance imaging measures of blood flow patterns in the human auditory cortex in response to sound. *Journal of Speech, Language, and Hearing Research*, 41, 538–48.

Jiang, J., Alwan, A., Keating, P., Auer, E. T., Jr., & Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing: Special issue on Joint Audio-Visual Speech Processing*, 2002, 1174–88.

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

Kaas, J. H. & Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 11793–9.

Knight, R. T., Scabini, D., Woods, D. L., & Clayworth, C. (1988). The effects of lesions of superior temporal gyrus and inferior parietal lobe on temporal and vertex components of the human AEP. *Electroencephalography and Clinical Neurophysiology*, 70, 499–509.

Krolak-Salmon, P., Henaff, M. A., Tallon-Baudry, C., Yvert, B., Fischer, C., Vighetto, A., Betrand, O., & Mauguiere, F. (2001). How fast can the human lateral geniculate nucleus and visual striate cortex see? *Society for Neuroscience Abstracts*, 27, 913.

Liberman, A. M. & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, 4, 187–96.

Liégeois-Chauvel, C., de Graaf, J. B., Laguitton, V., & Chauvel, P. (1999). Specialization of left auditory cortex for speech perception in man depends on temporal coding. *Cerebral Cortex*, 9, 484–96.

Liégeois-Chauvel, C., Musolino, A., Badier, J. M., Marquis, P., & Chauvel, P. (1994). Evoked potentials recorded from the auditory cortex in man: Evaluation and topography of the middle latency components. *Electroencephalography and Clinical Neurophysiology*, 92, 204–14.

Ludman, C. N., Summerfield, A. Q., Hall, D., Elliott, M., Foster, J., Hykin, J. L., Bowtell, R., & Morris, P. G. (2000). Lip-reading ability and patterns of cortical activation studied using fMRI. *British Journal of Audiology*, 34, 225–30.

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

Massaro, D. W. (1989). Multiple book review of *Speech Perception by Ear And Eye: A Paradigm for Psychological Inquiry*. *Behavioral and Brain Sciences*, 12, 741–54.

Massaro, D. W. (1999). Speechreading: Illusion or window into pattern recognition. *Trends in Cognitive Sciences*, 3, 310–17.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, 100, 1777–86.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–8.

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

Mesulam, M. M. (1994). Neurocognitive networks and selectively distributed processing. *Revue Neurologique*, 150, 564–9.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, 121, 1013–52.

Mitzdorf, U. (1986). The physiological causes of VEP: Current source density analysis of electrically and visually evoked potential. In R. Q. Cracco & I. Bodis-Wollner (eds.), *Evoked Potentials* (pp. 141–54). New York: Alan R. Liss Inc.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research*, 14, 115–28.

Mottonen, R., Krause, C. M., Tiippana, K., & Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research*, 13, 417–25.

Moutoussis, K. & Zeki, S. (1997). A direct demonstration of perceptual asynchrony in vision. *Proceedings of the Royal Society of London*, Series B: Biological Sciences, 264, 393–9.

Munhall, K. G. & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America*, 104, 530–9.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58, 351–62.

Näätänen, R. & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, 24, 375–425.

Nobre, A. C., Allison, T., & McCarthy, G. (1994). Word recognition in the human inferior temporal lobe. *Nature*, 372, 260–3.

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., Sensolo, S., & Fazio, F. (in press). A functional-anatomical model for lip-reading. *Journal of Cognitive Neuroscience*.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, 37, 127–52.

Ponton, C. W. & Eggermont, J. J. (2001). Of kittens and kids: Altered cortical maturation following profound deafness and cochlear implant use. *Audiology & Neuro-otology*, 6, 363–80.

Ponton, C. W., Don, M., Waring, M. D., Eggermont, J. J., & Masuda, A. (1993). Spatio-temporal source modeling of evoked potentials to acoustic and cochlear implant stimulation. *Electroencephalography and Clinical Neurophysiology*, 88, 478–93.

Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures: A functional magnetic resonance imaging study. *Journal of Neuroscience*, 16, 5205–15.

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, 18, 2188–99.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, 28, 617–25.

Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, 8, 516–21.

Rauschecker, J. P. & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 11800–6.

Rosenblum, L. D. (2002). The perceptual basis for audiovisual speech integration. *Proceedings of the 7th International Conference on Spoken Language Processing*, September 16–20, Denver, CO.

Saldaña, H. M. & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America*, 95, 3658–61.

Scherg, M. & Von Cramon, D. (1985). Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. *Electroencephalography and Clinical Neurophysiology*, 62, 32–44.

Scherg, M. & Von Cramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalography and Clinical Neurophysiology*, 65, 344–60.

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd, & D. Burnham (eds.), *Hearing by Eye II: The Psychology of Speechreading and Auditory-visual Speech* (pp. 85–108). Hove, UK: Psychology Press.

Scott, S. K. & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, 26, 100–7.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123, 2400–6.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, 59, 73–80.

Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, 90, 1797–1805.

Sekiyama, K. & Tohkura, Y. (1993). Interlanguage differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21, 427–44.

Sharma, A. & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America*, 106, 1078–83.

Sharma, A. & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America*, 107, 2697–703.

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *Journal of the Acoustical Society of America*, 108, 3030–5.

Singer, W. (1998). Consciousness and the structure of neuronal representations. *Philosophical Transactions of the Royal Society of London*, Series B: Biological Sciences, 353, 1829–40.

Stein, B. E. (1998). Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. *Experimental Brain Research*, 123, 124–35.

Stein, B. E. & Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, MA: MIT Press.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., & Howard, M. A., 3rd (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *Journal of Neurophysiology*, 82, 2346–57.

Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–15.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (eds.), *Hearing by Eye: The Psychology of Lip-Reading* (pp. 3–52). Hove, UK: Psychology Press.

Summerfield, Q. & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 36, 51–74.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive Neuropsychology*, 17, 257–79.

Trout, J. D. (2001). The biological basis of speech: What to infer from talking to the animals. *Psychological Review*, 108, 523–49.

Ungerleider, L. G. & Haxby, J. V. (1994). "What" and "where" in the human brain. *Current Opinion in Neurobiology*, 4, 157–65.

Vaughan, H. & Arezzo, J. (1988). The neural basis of event-related potentials. In T. W. Picton (ed.), *Human Event-Related Potentials* (pp. 45–96). Amsterdam: Elsevier Science Publishers.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, 57, 1124–33.

Watson, J. D., Myers, R., Frackowiak, R. S., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., Shipp, S., & Zeki, S. (1993). Area V5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex*, 3, 79–94.

Wise, R. J., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., & Warburton, E. A. (2001). Separate neural subsystems within "Wernicke's area." *Brain*, 124, 83–95.

Yehia, H., Kuratate, T., & Vatikiotis-Bateson, E. (1999). Using speech acoustics to drive facial motion. *Proceedings of the International Congress of Phonetic Sciences* (ICPhS 1999) (pp. 631–4). San Francisco, CA: IPA.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26, 23–43.

Yvert, B., Crouzeix, A., Bertrand, O., Seither-Preisler, A., & Pantev, C. (2001). Multiple supratemporal sources of magnetic and electric auditory evoked middle latency components in humans. *Cerebral Cortex*, 11, 411–23.

Zeki, S. (1998). Parallel processing, asynchronous perception, and a distributed system of consciousness in vision. *Neuroscientist*, 4, 365–72.

Zeki, S. (2001). Localization and globalization in conscious vision. *Annual Review of Neuroscience*, 24, 57–86.