Speech Perception

Cognitive Foundations and Cortical Implementation

David Poeppel[1,2] and Philip J. Monahan[1]

[1]Department of Linguistics, University of Maryland College Park, and [2]Department of Biology, University of Maryland College Park

ABSTRACT—*Speech perception includes, minimally, the set of computations that transform continuously varying acoustic signals into linguistic representations that can be used for subsequent processing. The auditory and motor subroutines of this complex perceptual process are executed in a network of brain areas organized in ventral and dorsal parallel pathways, performing sound-to-meaning and sound-to-motor mappings, respectively. Research on speech using neurobiological techniques argues against narrow motor or auditory theories. To account for the range of cognitive and neural attributes, integrative computational models seem promising.*

KEYWORDS—*functional anatomy of speech; dual-stream model; dorsal and ventral streams*

Understanding how speech signals are represented and processed in the human brain remains a core challenge for cognitive science and neuroscience. The neural basis of speech processing constitutes a fruitful area of inquiry at the interface between the cognitive and brain sciences insofar as there exist explicit and detailed models about the nature of the speech signal. First, there are hypotheses about what the elementary building blocks (or primitives) are. Second, there exists a growing body of data on the cortical regions and mechanisms that form the basis for processing speech.

DEFINITIONAL ISSUES

An issue requiring clarification at the outset is that the term *speech perception* is used in varied contexts. Importantly, it does not refer to language comprehension in general but only to one subroutine of comprehension. Comprehension is a set of linguistic computations that can be initiated by auditory (speech), visual (text or sign), or somatosensory input (Braille). In contrast, speech perception refers to the set of operations that transform an auditory signal into mental representations of a type that can make contact with internally stored information—that is, words.

There are multiple levels of representation to consider in the mapping from auditory signal to perceptual interpretation (Fig. 1). It is therefore critical to distinguish among the uses of the term speech perception. Historically, most work focused on the perception of single speech sounds (phones or phonemes or segments, labels that have noninterchangeable, precise, technical meanings) or syllables (e.g., consonant-vowel, or CV, syllables). This early work thus investigated sublexical aspects of speech. Phonemes, or segments, are not the smallest hypothesized units of representation, however. Consequently, many experiments tested the role of "distinctive features," the putatively smallest units for the representation of speech sounds, most often stated as *articulatory primitives*—for example, [±coronal] (i.e., whether or not the tongue blade is implicated in the production of a given sound) or [± voiced] (i.e., whether or not the vocal folds vibrate during production of a given sound). Experiments focusing on the sublexical properties of speech are typically associated with phenomena such as categorical perception, prototype effects, assimilation, and related metrics.

Unsurprisingly, other studies approach perceptual issues from the perspective of recognizing spoken words. Here, the commitment to the format of lexical representation becomes a central issue: If words are represented using an abstract code (such as a featural one), then the mapping from sound to that code forces a set of computations that translate the acoustic input into a discretized representation. In contrast, if words are represented as, say, acoustic exemplars, then the mapping is strikingly different, and little computation is required to translate the acoustic input into the hypothesized representation (see Poeppel, Idsardi, & van Wassenhove, 2008, for discussion). Studies of spoken-word recognition obviously emphasize word-level tasks—for example, lexical decision (word/non-word judgments)—to test semantic or phonological relations between the items in one's mental lexicon.

Another prominent and growing body of recent research in cognitive neuroscience approaches speech recognition from a rather different, sentence-level, perspective, dealing principally

Address correspondence to David Poeppel, Department of Linguistics & Department of Biology, 1401 Marie Mount Hall, University of Maryland College Park, College Park, MD 20742; e-mail: dpoeppel@umd.edu.

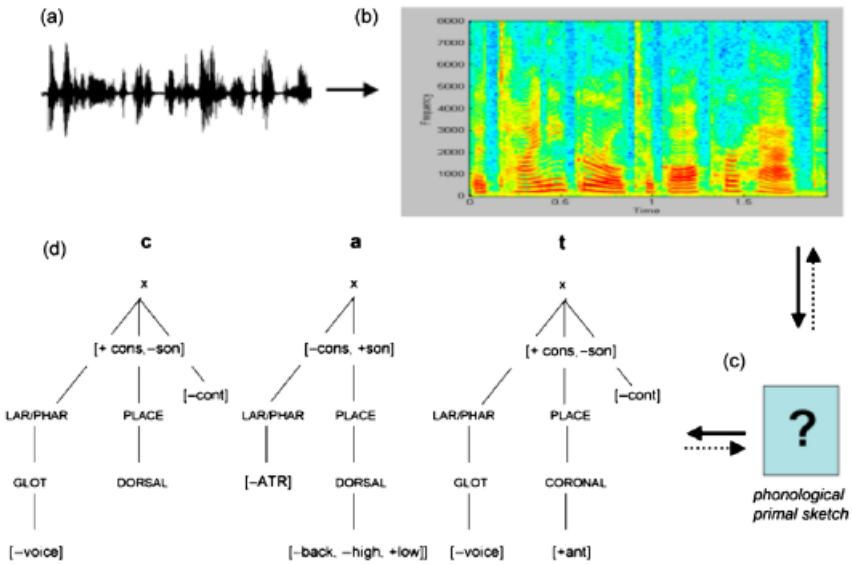

Fig. 1. Representations and transformations from auditory signal to lexical representation. At the periphery, the listener encodes a continuously varying waveform (a). The afferent auditory pathway analyzes the input signal in the time and frequency domains. A neural version of the spectrogram (b) is generated, highlighting both spectral and temporal variation over multiple time scales. An intermediate representation (c), here called ''phonological primal sketch,'' may be essential to map from an acoustic to a putatively abstract representation of auditory signals. The phonological primal sketch might comprise temporal primitives (e.g., temporal integration windows, or slices of time over which the brain extracts and integrates information of specific durations) and spectral primitives (e.g., combinations of frequencies to which the auditory brain reacts in a privileged manner). The diagram (d) shows how the word *cat* may be represented in the mind/brain of the speaker/listener. Each of the three segments of the consonant-vowel-consonant (CVC) syllable is assembled from distinctive features, the hypothesized primitives that are the smallest units of speech and have both articulatory (e.g., [±coronal]) and acoustic (e.g., [± sonorant]) interpretations.

with the concept of "intelligibility." In these studies, participants are presented with spoken sentences (often manipulated to vary some parameter under investigation) and asked to report what they heard, often while brain activity is monitored (Luo & Poeppel, 2007; Scott, Blank, Rosen, & Wise, 2000). Naturally, subjects execute the task demands by accessing speech representations at all levels, including syntactic, lexical, morphemic, syllabic, and featural information. Thus, while intelligibility studies at the sentence level have desirable properties with respect to ecological validity, they can be more difficult to interpret due to the fact that speech-based representations participate at all levels.

Given the intricacy of the perceptual challenge and the variety of representations relevant to recognition, it stands to reason that the neuronal basis is complex. There is, for example, no single cortical region that can be argued to be principally responsible for speech perception (although the superior temporal sulcus appears to mediate some critical computations). This is unlike face-recognition research, where one particular cortical field—the fusiform face area—has been argued to play a disproportionately large role (but see Grill-Spector and Sayres, 2008, this issue). The cognitive sciences have implicated a range of computational subroutines; similarly, data from neuropsychology, neuroimaging, and electrophysiology converge on the view that speech perception is mediated by a network of interconnected regions in the frontal, parietal, and temporal lobes, with the different areas making specific, task-modulated contributions in the mapping from sound to meaning and from sound to articulatory representation.

A DUAL-STREAM MODEL

Contemporary approaches to the brain basis of speech reflect two paradigmatic shifts. First, research on the acoustics of speech has shifted emphasis from more spectrally based investigations (e.g., what is the difference between /ba/ and /ga/ in frequency space?) to more temporally based approaches, in which concepts such as the *envelope* of a speech signal (i.e., fluctuations in power over time) or the *modulation spectrum* (i.e., which sound frequencies are modulated at which rates) are used to account for perceptual phenomena (Greenberg & Ainsworth, 2006). One might summarize this shift as changing emphasis from single sounds to connected speech and from spectral to temporal modulation. A second perspective shift concerns the potent role that task demands play. It is now uncontroversial that the functional anatomy of speech perception is strongly conditioned by the

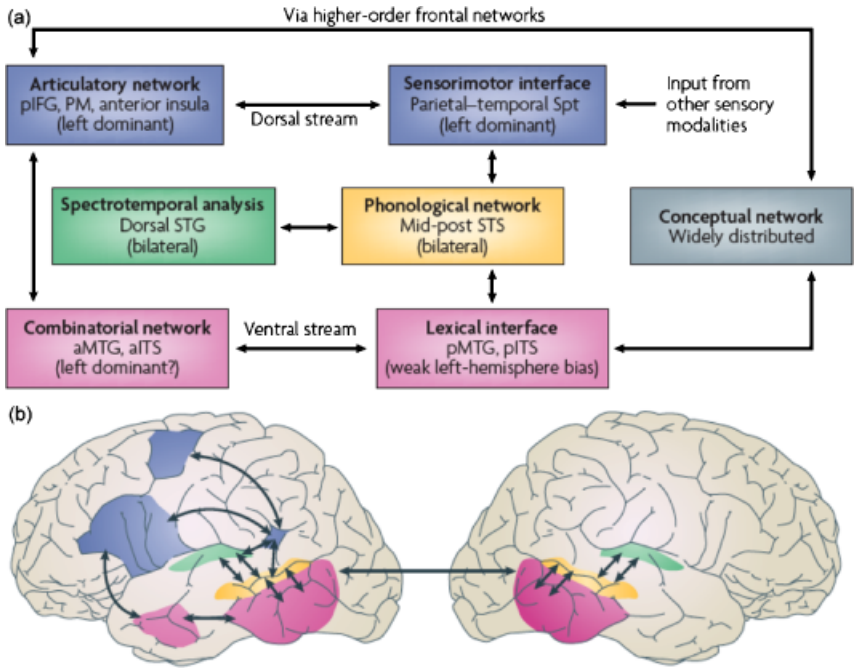

Fig. 2. Dual-stream model of the functional anatomy of speech perception and language comprehension. The first stage of cortical speech processing involves a spectrotemporal analysis (green box), arguably carried out in the core and belt areas of the superior temporal cortex, bilaterally. The analyses carried out in the left and right auditory regions appear to differ: The propensity for analyzing lower modulation frequencies (longer temporal integration windows commensurate with the analysis of suprasegmental information) is more strongly developed in the right hemisphere. The ''phonological network'' implicates the middle and posterior aspects of the superior temporal sulcus (STS) bilaterally, although possibly with a left-hemisphere bias. Subsequently, the system diverges into two broad streams, a dorsal pathway (blue boxes) that maps auditory/phonological representations onto articulatory/motor representations, and a ventral pathway (purple boxes) that maps phonological representations onto lexical conceptual representations. Approximate locations of the cortical regions in the dual-stream model appear in the brain diagrams at bottom. (aITS = anterior inferior temporal sulcus; a(p)MTG = anterior (posterior) middle temporal gyrus; pIFG = posterior frontal gyrus; PM = premotor cortex; Spt = Sylvian parietotemporal area.) Reprinted from ''The Cortical Organization of Speech Processing,'' by G. Hickok and D. Poeppel, 2007, *Nature Reviews Neuroscience*, *8*, p. 395. Copyright 2007, Nature Publishing Group. Reprinted with permission.

perceptual "endgame." For example, computational subroutines that mediate making contact with lexical representations are associated with temporal lobe systems; in contrast, when it is critical to recruit the articulatory representations underlying speech, parietal and frontal cortical fields are implicated.

One relatively generic model that attempts to capture these recent developments and integrate the cognitive requirements of speech perception with known neuropsychological and neuroimaging findings postulates that there is a dual stream of information processing (Hickok & Poeppel, 2007), as illustrated in Figure 2. The incoming signal's spectrotemporal properties are initially analyzed in the dorsal and posterior superior temporal gyrus (STG) and superior temporal sulcus (STS). Critically, these early computations are mediated bilaterally in the superior temporal cortex (Binder et al., 2000), although the left and right cortical areas have important computational specializations (with regard to timing properties) that contribute differentially to the recognition process (Hickok & Poeppel, 2007; Poeppel et al., in press).

Two processing streams originate from this early spectrotemporal analysis. A *ventral pathway* incorporates middle temporal gyrus, inferior temporal sulcus, and perhaps the inferior temporal gyrus. The ventral stream maps from sensory/phonological representations to lexical or conceptual representations (i.e., sound to meaning). A *dorsal pathway*, including the Sylvian parietotemporal area (SPT) as well as the inferior frontal gyrus, anterior insula, and premotor cortex, forms the substrate for mapping from sensory/phonological representations to articulatory-motor representations. While early cortical analysis is indisputably bilateral and much of the processing in the ventral stream is more bilateral than previously assumed (Binder et al., 2000; Hickok & Poeppel, 2007), the dorsal pathway is left-lateralized. Evidence that supports such an analysis derives from neuropsychological deficit-lesion data, hemodynamic

imaging data, and electrophysiological data (electroencephalography [EEG], magnetoencephalography [MEG]).

Functional anatomic models of this type might begin to meet the challenges posed by the "granularity mismatch" existing between cognitive-science theories and neurobiological practice (i.e., the inability to explicitly link representations and mechanisms across disciplines due to positing explanations at different levels of analysis). For now, it remains entirely unclear how the putative primitives of speech (e.g. feature, syllable, etc.) map onto the putative primitives of the biological substrate (e.g., neuron, synapse, oscillation, etc.). If we can obtain some understanding of the computational contribution of individual cortical areas to the perceptual process, we can impose compelling constraints on cognitive models. Additionally, cognitive science supplies the primitive representations and operations that, if spelled out in computational terms at the appropriate level of abstraction, can stimulate new ideas about neurobiological implementation. In short, linking hypotheses between speech and brain are most likely to bear fruit if they make use of computational analyses that appeal to generic computational subroutines.

FOUR RESEARCH DOMAINS: SOME TOPICS WITH LEGS

Here we point to recent studies that continue to stimulate new perspectives on how the cognitive sciences and neurosciences must interact in a nuanced manner to generate theoretically motivated, computationally explicit, and biologically realistic approaches to speech processing. These represent areas of inquiry that, in our view, will shape the debates on how speech is represented and processed in the human brain.

Abstraction

Cognitive neuroscience data have been used to weigh in on debates regarding the nature of linguistic representations. Linguists traditionally use native speakers' intuitions and cross-linguistic observations as evidence in support of psychologically abstract linguistic representations. Electrophysiology is now proving to be an effective tool in elucidating representational questions and lending support to theoretically motivated claims. The evidence accumulated thus far is suggestive: Abstract linguistic representations constrain our processing of the speech signal.

In mismatch-negativity (MMN in EEG; mismatch magnetic field, MMF, in MEG) studies, participants passively listen to a series of auditory stimuli, many of which are either identical or share some relevant property. Occasionally, a stimulus that is either entirely different from the "standard" or differs on some congruous attribute is presented. The amplitude of the response to this "deviant" is larger than the attenuated response to the standard. Importantly, the main neural generator of the MMN/MMF lies in the superior temporal cortex. Effects observed in such designs must thus be attributed, at least in substantial part, to the auditory cortex.

Näätänen et al. (1997) studied the effects of native language vowel inventory on early auditory processing. Specifically, they asked whether native language representations constrained the early auditory processing of vowels. Both Finnish and Estonian speakers heard /e/ as the standard and /ö/ and /õ/ as the deviants. The vowel systems of both Finnish and Estonian contain /ö/, but only Estonian has /õ/. This is the only difference between the vowel systems of the two languages. An MMN was elicited in both Finnish and Estonian speakers when /ö/ was the deviant. Crucially, however, when /õ/ was the deviant, an MMN was elicited only in Estonian speakers and not in Finnish speakers. Acoustically, /õ/ and /ö/ are equally complex and equally distinct from /e/. Acoustic differences alone are, therefore, insufficient to account for why the MMN was elicited in Estonian and not Finnish participants. Instead, native language representations appear to constrain early auditory processing.

Kazanina, Phillips, & Idsardi (2006) asked a similar question. They tested Russian and Korean participants on the /t/ and /d/ contrast. Russian speakers can perceptually discriminate these two sounds, while Korean speakers cannot. Accordingly, it is claimed that the abstract inventory of sounds in Russian contains both /t/ and /d/, while the inventory in Korean contains only one sound, /T/. Both groups of participants heard one block in which /t/ was the deviant and /d/ was the standard and another block reversing standard and deviant. Like the Estonians in the Näätänen et al. (1997) study, a large MMF was elicited in Russian speakers for the deviant. No MMF was elicited in the Korean participants. This effect cannot solely be attributed to acoustic differences, given that an MMF was found in Russian and not Korean participants. Instead, native-language linguistic representations constrain early auditory processing. As these findings suggest, electrophysiological techniques can be used to adjudicate between claims regarding the nature of linguistic representations.

Sound–Motor Mapping

Recent imaging studies have revived questions regarding the mapping between acoustic information and articulatory representations that underlies the generation of speech. One influential strand of research, motivated by the motor theory, has long maintained that access to (aspects of the) production mechanisms—perhaps in the context of internal forward models—is critical to the successful perceptual analysis of speech. In contrast, recent work has championed auditory theories, demonstrating that the sophisticated machinery of the auditory pathway permits extracting richly structured information from complex input signals (Greenberg & Ainsworth, 2006).

Experiments using fMRI have now shown (in auditory, as well as auditory-visual conditions; see below) that cortical regions canonically implicated in motor tasks are recruited for perception. For example, Wilson, Saygin, Sereno, & Iacoboni (2004)

show that motor areas (in the frontal cortex) are robustly activated in basic speech tasks. Moreover, Hickok, Buchsbaum, Humphries, & Muftuler (2003) have identified an area at the juncture of the temporal and parietal lobes (area SPT; see Fig. 2) that may be the cortical substrate that enables coordinate transformations from acoustic to motor coordinates. Given, on the one hand, data showing motor cortex involvement in perception and, on the other, the ability of listeners to extract interpretable information from synthesized or artificially modified speech that is not closely related to vocal-tract generation, one can surmise that "auditory only" or "motor only" theories will not suffice. Instead, the data suggest that the perceptual problem is so complex that any relevant information source is used.

Audiovisual speech

Cognitive neuroscience also effectively informs psychological models of perception in audiovisual speech integration. Here, too, models espousing abstract representations that constrain lower-level perceptual processes are finding support. For example, using EEG, van Wassenhove, Grant, & Poeppel (2005) found that the degree of ambiguity in visual speech predicts the speed at which an auditory speech signal is processed. The more transparent the visual signal (viseme) is, the faster the auditory signal is processed in auditory cortex. This is in line with models that predict that the use of higher-order information facilitates or constrains early auditory processing. Skipper, van Wassenhove, Nusbaum, & Small (2007), using fMRI, found that many of the same cortical areas involved during speech-production tasks were implicated in audio-visual perception. Specifically, they concluded that listeners create a motor plan of the intended utterance and that this motor plan influences the percept. These findings provide additional evidence that the cortical mechanisms underlying perception and production are shared, a point amplified below. Van Wassenhove et al. (2005) and Skipper et al. (2007) interpret their findings as supporting *analysis-by-synthesis* models of perception. In these models, top-down hypotheses are generated based on the available information and these hypotheses modulate lower-level analyses. Neuroimaging data on high-level vision have also revealed a modulatory role of top-down prediction in shaping visual cortical responses during object perception (e.g., Bar, 2007), and the notion of analysis-by-synthesis has similarly received support (e.g., Yuille & Kersten, 2006).

Speech Production

While we focus on speech perception, research on production underscores the tight link between perceptual and productive mechanisms. An extensive meta-analysis of imaging data on speech production by Indefrey & Levelt (2004), for example, shows extensive overlap between the cortical networks for production with those for perception. The neurophysiological mechanisms that form the basis for speech processing are not segregated for perception and production processes but, instead, draw on common brain areas that execute computations germane to both processes. Imaging experiments by Guenther, Ghosh, & Tourville (2006) exploring the interface of computation and neurobiology support a production model that thoughtfully captures the relation between production mechanisms and the internal forward model that plays a key role in perceptual analysis. This research demonstrates the promise of applying computational models to develop linking hypotheses between cortical substrate and cognitive process.

PROSPECTS

We emphasized how neurobiological data from a range of approaches place constraints on psychological models of speech processing. However, at least historically, the flow of information has typically been from cognitive model to neurobiological experiment; it is only recently that biological data play a core role in shaping theories of speech. For example, an enormous amount of research was stimulated by the motor theory of speech perception (Liberman & Mattingly, 1985), which posits (intended) articulatory gestures as the representational primitives in speech. Countless neuroimaging studies have tested motor-theoretic claims, and at least some tight mapping between perception and production must be acknowledged (see above). A contrasting perspective builds on the assumption that speech is, principally, an auditory phenomenon, and that acoustic primitives must be recovered from the signal. In the context of this tradition, exemplified by Stevens (2002) as well as the research summarized in Greenberg & Ainsworth (2006), neurophysiological studies seek to identify the neural correlates of the putatively elementary auditory representations.

Further stimulating work tackling the mapping from cognition to neuronal implementation includes studies elucidating where in the auditory cortex acoustic speech features are processed—that is, a "spatial code" or "spatial map" for speech sounds (Obleser, Lahiri, & Eulitz, 2003)—and experimentation on how contextual information influences perceptual tasks (Friedrich & Kotz, 2007). Investigating the cortical basis of speech-sound processing is a fertile domain in which to explore how concepts from the cognitive sciences link mechanistically to biological mechanisms—in other words, to figure out the mapping from physics (vibrations in the ear) to cognition (abstractions in the head). Particularly promising are the data showing a tight relation between perception and production mechanisms, opening the path for using computational theories that build on internal forward models, predictive coding, and analysis-by-synthesis.

Recommended Reading

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, *8*, 389–395. Discusses how some elementary

acoustic attributes of signals, such as their temporal properties, contribute to functional asymmetries at the basis of speech sound processing.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*, 67–99. Provides an in-depth review of the ventral/dorsal pathway model.

Jacquemot, C., & Scott, S.K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*, 480–486. A model proposing that phonological short-term memory arises out of conversions between the speech-perception and speech-production systems.

Vouloumanos, A., & Werker, J.F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*, 159–164. Evidence concerning the debate on whether or not speech is "special."

Wong, P.C.M., Skoe, E., Russo, N.M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*, 420–422. Data suggesting that the brainstem may play a more central role in shaping cortical processing and reflect experience-dependent tuning of the afferent auditory pathway.

Acknowledgments—DP and PJM are supported by 2R01DC05660.

REFERENCES

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*, 280–289.

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*, 512–528.

Friedrich, C.K., & Kotz, S.A. (2007). Event-related potential evidence of form and meaning coding during online speech recognition. *Journal of Cognitive Neuroscience*, *19*, 594–604.

Greenberg, S., & Ainsworth, W.A. (2006). *Listening to speech: An auditory perspective*. Mahwah, NJ: Erlbaum.

Grill-Spector, K., & Sayres, R. (2008). Object recognition: Insights from advances in fMRI methods. *Current Directions in Psychological Science*, *17*, 73–79.

Guenther, F.H., Ghosh, S.S., & Tourville, J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*, 280–301.

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: Speech, music, and working memory in area SPT. *Journal of Cognitive Neuroscience*, *15*, 673–682.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402.

Indefrey, P., & Levelt, W.J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*, 101–144.

Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences, USA*, *103*, 11381–11386.

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

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*, 1001–1010.

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

Obleser, J., Lahiri, A., & Eulitz, C. (2003). Auditory-evoked magnetic field codes place of articulation in timing and topography around 100 milliseconds post syllable onset. *Neuroimage*, *20*, 1839–1847.

Poeppel, D., Idsardi, W., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *363*, 1071–1086.

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

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

Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

van Wassenhove, V., Grant, K., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, USA*, *102*, 1181–1186.

Wilson, S.M., Saygin, A.P., Sereno, M.I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*, 701–702.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–308.