Behavioral and Neuroimaging Approaches to ‘Speech Perception’

The term ‘speech perception’ is used in a variety of contexts. Critical terminological distinctions should be made. To begin, speech perception is not equivalent to language comprehension; rather, the perception of the speech signal constitutes only one subroutine of comprehension. The latter is comprised of a set of linguistic computations that can be initiated by auditory input (speech), visual input (text or sign), or somatosensory input (Braille). ‘Speech perception’ thus refers to the set of operations that transform an auditory signal into representations of a form that makes contact with internally stored information – that is, the stored words in a listener’s mental lexicon.

There are at least three empirical approaches that are typically grouped under the term. Because they are different in terms of the structure of the input signal, the perceptual subroutines implicated, and the endpoint of the processes, it is important to be aware of these distinctions.

(a) A large body of research on speech perception addresses effects at the level of individual speech sounds, that is, sublexical or prelexical units of speech. Subjects are typically exposed to vowels or syllables and execute particular behavioral tasks, such as discrimination or identification. For example, subjects listen to consonant–vowel (CV) syllables constructed from an acoustic continuum – for example, a series exemplifying a /ra/-/la/ tongue-shape contrast or a /bi/-/pi/ voicing contrast or a /tu/-/ku/ place-of-articulation contrast – and asked on presentation of a single token to identify the category. Such designs focus on sublexical properties of speech and examine questions concerning, for example, the nature of categorical perception in speech (e.g., Liberman, 1996), the phonemic inventory of speakers/listeners of different languages (e.g., Harnsberger, 2000), perceptual magnet effects (e.g., Kuhl et al., 2008), changes associated with first (e.g., Eimas, Siqueland, Jusczyk, & Vigorito, 1971) and second language learning (e.g., Flege & Hillenbrand, 1986), phonotactic constraints (e.g., Dupoux, Kakahi, Hirose, Pallier, & Mehler, 1999; Kabak & Idsardi, 2007), the role of distinctive features (e.g., Kingston, 2003), and other issues that can be experimentally evaluated at the prelexical level of analysis.

This research strategy has been highly productive in the behavioral literature, and experiments of this type are now also prominent in the cognitive neurosciences, both in the context of imaging (fMRI, PET) and electrophysiological approaches (EEG, MEG, ECog). For example, using fMRI, several teams have examined regionally specific hemodynamic effects when subjects perform judgments on categorically varying stimuli (Blumstein, Myers, & Rissman, 2005; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Raizada & Poldrack, 2007). These studies aim to identify areas responding differentially to signals belonging to different categories or that differentiate speech versus nonspeech tokens (Binder et al., 2000). Unfortunately, no simple answer has resulted from even similarly designed studies, with temporal, parietal, and frontal areas implicated. One emerging generalization is that bilateral posterior superior temporal gyrus (pSTG) is critical in phonemic analysis, although the wide range of activated areas across studies prevents a definitive conclusion. Electrophysiological methods have been used successfully with such sublexical materials, for example, to probe the phonemic inventories of speakers of different languages. For example, Näätänen et al. (1997), using EEG and MEG, were able to demonstrate neurophysiological response profiles (arising from superior temporal cortex) that characterize the vowel inventories of Finnish versus Estonian speakers. Similarly, Kazanina, Phillips, and Idsardi (2006) used MEG data to illustrate how language-specific contrasts (Russian vs. Korean), including allophonic distinctions, can be quantified neurophysiologically. Recent data capitalizing on the high resolution that can be achieved with electrocorticography (ECoG) suggest that pSTG reflects rapid categorical phonetic analysis (Chang et al., 2010).

While this research approach has had considerable influence, it must be acknowledged that there are important limitations. For example, a large number of studies investigate categorical perception as well as the notion of ‘rapid temporal processing,’ typically based on plosive contrasts. Syllables with plosive onsets are admirably fascinating in their complexity (and, e.g., voice onset time is easily manipulated), but a variety of other equally relevant phenomena at the prelexical level have not been explored. Moreover, the studies under discussion are ecologically completely invalid and therefore hard to generalize: experimenters present single, sublexical pieces of speech in the context of experimental settings that require selective attention to particular features and engage no further linguistic processing. The results are in danger of either masking or distorting the processes of interest that are responsible for ecologically more natural speech perception. In canonical perceptual contexts, speakers/listeners do not consciously attend to sublexical material; therefore, the interpretation of the results requires caution, as task demands are known to modulate neuronal reactivity in striking ways.

(b) A different approach investigates speech perception from the perspective of spoken word recognition. Various lexical access models derive from such research (for instance, lexical access from spectra, Klatt, 1979, 1985; instantiations of the cohort model, e.g., Gareth Gaskell & Marslen-Wilson, 2002; neighborhood activation model, Luce & Pisoni, 1998; continuous mapping models, Alloppena, Magnuson, & Tanenhaus, 1998) and have yielded foundational information about the mental/physical representations that underpin stored words and their processing. Typical experimental manipulations in the context of speech perception include lexical decision, naming, and priming. Recognizing single, spoken words is a more natural task than performing unusual experimental demands...
on sublexical material such as single vowels or syllables. Some models of speech perception and lexical processing, such as the TRACE model (McClelland & Elman, 1986), view featural and lexical access as fully integrated whereas others argue for more cascaded operations.

Important empirical contributions to the neural basis of spoken word recognition have been made in the imaging literature, notably by Blumstein and colleagues. They have examined the issues using both lesion and imaging data (e.g., Misiurski, Blumstein, Rissman, & Berman, 2005; Prabhakaran, Blumstein, Myers, Hutchison, & Britton, 2006; Utman, Blumstein, & Sullivan, 2001). The data they have published support a model in which superior temporal areas mediate acoustic–phonetic analyses, temporoparietal areas perform a mapping to phonological–lexical representations, and frontal areas (specifically the inferior frontal gyrus) are integral to resolving competition between alternatives when listeners are presented noisy or underspecified input signals. The effect of lexical status on speech sound categorization has been studied in the behavioral literature (often to assess top-down effects, i.e., does knowing a word shift perceptual boundaries?), and Blumstein and colleagues, using voicing continua with word or nonword endpoints, have extended this work using fMRI (e.g., Myres & Blumstein, 2008). They demonstrate with fMRI dissociations between functionally early effects (in the temporal lobes: presumably related to perceptual aspects) and later, downstream (decision) processes implicating frontal lobe structures. Another powerful behavioral task in studies of lexical representation is repetition priming, and Gagnepain et al. (2008) used word and nonword repetition priming to determine which brain structures are sensitive to the activation of lexical entries. Bilateral superior temporal sulcus and superior temporal gyrus (STS, STG) were especially robustly implicated. Cumulatively, such data suggest that the mapping from acoustic–phonetic to lexical information occurs in cortical regions that are slightly more ventral than perceptual computations (and, crucially, occur bilaterally; cf. Hickok & Poeppel, 2000, 2004, 2007). The timing of lexical access has been studied using various electrophysiological approaches. Recent data using a repetition-priming design in the context of MEG recording suggest that around ~200 ms postword onset, lexical access is likely to have been achieved (Almeida & Poeppel, 2013). Importantly, there is a growing body of work examining the time course of lexical access, and while there is disagreement (on theoretical and methodological grounds), there is emerging consensus that by 150–250 ms, contact with lexical representations has been established and that the robust lexical effects between 300 and 400 ms reflect higher order features of lexical processing such as the calculation of neighborhood densities or surface frequencies. Electrophysiological approaches have also been used to test detailed theoretical proposals about lexical representation and processing. For example, Gagnepain, Henson, and Davis (2012) have tested ideas about predictive coding in lexical access, as reflected in auditory cortex; and Eulitz and colleagues (e.g., Friedrich, Eulitz, & Lahiri, 2006) have used lexical decision designs to assess the predictions arising from linguistic underspecification models of lexical representation. In sum, lexical-level experiments combined with neuroimaging tools have provided a fruitful arena linking linguistics, psychology, and neuroscience.

(c) A third approach to which the expression ‘speech perception’ is often applied is in the context of recognizing spoken sentences and assessing their intelligibility. In such experiments, participants hear sentences (containing various acoustic manipulations) and provide judgments of intelligibility, for example, by typing what they heard or reporting keywords or providing numerical metrics that reflect performance. Understanding how listeners process spoken sentences is an essential goal because it is the ecologically valid task we most want to explain. However, there is a price to pay for using naturalistic materials. Experimentally, it becomes challenging to isolate input-related, perceptual processes per se, since the presentation of sentences necessarily entails lexical processes, syntactic processes, both lexical semantic and compositional semantic operations – and therefore engages numerous ‘top-down’ processing factors that are known to play a critical role in the analysis of spoken input.

Sentence-level intelligibility studies – employing many manipulations of the acoustic structure of sentences – have been used in a number of neuroimaging and electrophysiological studies. In a series of PET and fMRI studies, for example, Scott and colleagues have shown that anterior temporal lobe structures, especially anterior STS, play a key role for intelligibility (e.g., Scott, Blank, Rosen, & Wise, 2000). Okada et al. (2010), in contrast, argued that more posterior regions are responsive to intelligible speech. Electrophysiological techniques have also been used to study sentence-level speech intelligibility, and Luo and Poeppel (2007), Doelling, Arnal, Glitzia, and Poeppel (2013), and Peelle, Gross, and Davis (2013) have argued that phase information in the cortical signal of a particular frequency, the theta band, is particularly closely related to and modulated by the acoustics of sentences. Recent work on the well-known cocktail party phenomenon has used both MEG and ECog to elucidate which features of neural responses track successful perceptual analysis when listeners are presented with complex listening situations with competing streams (e.g., Zion-Golumbic, Cogan, Schroeder, & Poeppel, 2007; Doelling, Arnal, Glitzia, and Poeppel, 2013, and Peelle, Gross, and Davis, 2013).

The locution ‘speech perception’ is used in the literature in at least three different ways. Important features of the neurocognitive systems that underpin speech and language have been discovered using all three approaches. This above brief summary serves to caution the reader that it is challenging to isolate relevant perceptual computations. Functionally speaking, speech perception comprises the set of operations that take as input continuously varying acoustic waveforms made available at the auditory periphery and that generate as output those representations (phonemic, syllabic, morphemic, lexical) that serve as the data structures for subsequent computations that underlie language comprehension.

**Integrative Proposals: Dual Streams**

Recent approaches to the brain basis of speech perception reflect two major shifts in research emphasis. Basic research on the acoustics of speech has shifted focus a bit, from more spectrally oriented research (e.g., what are the spectrotemporal differences between /ba/ and /ga/ and how is such a difference encoded) to temporally based approaches that highlight features such as the ‘envelope’ of a speech signal or the ‘modulation spectrum,’ both now used to account for numerous perceptual phenomena. This shift reflects changing emphasis...
from single speech sounds to connected speech as well as from spectral to temporal modulation. A second perspectival shift comes from the surprisingly rich role that task demands play in neuroimaging experiments. It is well established that the functional anatomy of speech perception is very clearly conditioned by the perceptual goals that are imposed by the task demands. For example, subroutines that underpin making contact with lexical representations are associated with temporal lobe systems; in contrast, when tasks require the recruitment of articulatory representations underlying speech, parietal as well as frontal regions are implicated.

One large-scale model that aims to capture recent developments and integrate the cognitive requirements of speech perception with neuropsychological and neuroimaging findings postulates that there is a dual stream of information processing (Hickok & Poeppel, 2007), as illustrated in Figure 1. A proposal that is similar in spirit – albeit different in many of relevant details – is Rauschecker and Scott’s (2009) model. The HP 2007 model states that an incoming signal’s spectrotemporal properties are analyzed in the dorsal and posterior STG and STS. Importantly, early computations are mediated bilaterally, in the superior temporal cortex (as demonstrated well by fMRI data by Binder et al., 2000), although the left and right cortical areas have important computational specializations that contribute differentially to the recognition process (Hickok & Poeppel, 2007).

By hypothesis, two processing streams originate from this early spectrotemporal analysis. A ventral pathway (and more likely more than one projection) extends to STG and incorporates middle temporal gyrus (MTG), inferior temporal sulcus, and perhaps the inferior temporal gyrus. There is debate about the importance of more anterior versus posterior aspects of the ventral stream(s). Functionally, the ventral stream maps from sensory/phonological representations to lexical or conceptual representations (i.e., mapping from ‘sound’ to ‘meaning’). A

![Figure 1](Image)
dorsal pathway (also more likely more than one projection), 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 had previously been argued (Binder et al., 2000; Hickok & Poeppel, 2000, 2004, 2007), the dorsal pathway has been argued to be more left-lateralized. Evidence that supports such an analysis derives from neuropsychological deficit-lesion data, hemodynamic imaging data, and electrophysiological data (EEG, MEG). However, recent ECog data challenge the extent to which the dorsal projections that underlie sensorimotor transformations are left-lateralized (Cogan et al., 2013). It is worth noting that dual stream models have also been extended to lexical-level processing (e.g., Gow, 2012) and sentence-level processing (e.g., Friederici, 2012), and there is emerging consensus that such a conceptualization of the functional anatomic architecture of speech and language processing provides a productive research agenda.

**Neuroimaging of Speech Perception: Current Areas of Active Debate**

Several research areas (across levels of processing, i.e., sounds, words, and sentences; and across techniques, i.e., fMRI, MEG, ECog) are receiving a fair amount of attention and generating controversy.

(a) From the perspective of functional anatomy, the issue of lateralization of speech perception has generated a large body of empirical work and yielded robust controversy. Historically, based primarily on deficit-lesion data, speech perception – and language processing more generally – has been argued to be a strictly left-lateralized phenomenon. Neuroimaging as well as neuropsychological findings challenge that generalization and support a bilateral set of neural structures underpinning speech perception (Binder et al., 2000; Poeppel et al., 2004). Nevertheless, there exist striking asymmetries – and the origin of the potential asymmetries has fostered serious debate. Broadly speaking, some researchers maintain that lateralization is functionally driven (McGettigan & Scott, 2012); other researchers argue that the two hemispheres execute slightly different forms of computational analyses (Poeppel, 2003; Zatorre, Belin, & Penhune, 2002), based on imaging data (Boenio, Fromm, Braun, & Poeppel, 2005; Olesen, Eisner, & Kotz, 2008).

(b) From the perspective of stimulus manipulations that can be used to investigate speech perception in novel ways (behaviorally and neurally) two directions have proven productive: the processing of audiovisual speech and the use of different signal distortions to assess perception. (i) A range of signal distortions has been used to interrogate the system. Some manipulations focus on the spectral content of speech (e.g., noise vocoding, e.g., Peelle, Gross, & Davis, 2013; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; sine wave speech, e.g., Möttönen et al., 2006; Remez, Rubin, Pisoni, & Carrell, 1981), while other approaches play with the temporal structure (e.g., compression, e.g., Ahissar et al., 2001; local reversal, e.g., Saberi & Perrott, 1999). Such experiments typically assess the tolerance to spectral and temporal distortions of speech signals in listeners and then measure which brain regions reflect sensitivity to such distortions or to signal intelligibility. Superior temporal cortical areas are modulated by the spectral and temporal features of the signal, but which regions are differentially sensitive to spectral or temporal features of speech remains under debate. However, speech intelligibility is largely reflected in the activation of the STS, although both anterior and posterior aspects have been implicated in different studies.

(ii) Audiovisual speech studies have been popular, across recording techniques. With respect to the anatomic locus of audiovisual speech integration, the imaging evidence is converging that posterior STS provides the substrate for the binding of the auditory and visual information to underpin multisensory speech (Nath & Beauchamp, 2011; Okada, Venezia, Matchin, Saberi, & Hickok, 2013). STS receives input from the appropriate sensory regions and is situated to interface with the face and language processing regions identified to be critical (cf. Figure 1). A number of imaging and electrophysiology studies have attempted to dissect functionally the operations during AV speech. van Wassenhove, Grant, and Poeppel (2005) and Arnal, Morillon, Kell, and Giraud (2009) found, using EEG and MEG, that the degree of ambiguity in visual speech predicted the speed at which the corresponding auditory signals were processed: the more transparent the visual signal (viseme) is, the faster the auditory signal is processed in auditory cortex. This aligns with models that predict that the use of higher order information facilitates early auditory processing. van Wassenhove et al. (2005) interpret their findings as supporting analysis-by-synthesis models of perception. In such models, top-down hypotheses are generated based on the available sparse information and the hypotheses constrain lower level analyses.

(c) Some of the hypothesized computations underlying speech perception are being addressed in cognitive neuroscience studies. For example, how is the continuously varying speech input parsed into units of the appropriate granularity for further processing (the segmentation problem)? Recent neuropsychological research using EEG, MEG, and intracranial data suggests that slow cortical oscillations (<10 Hz) may play a key role in chunking the stream into usable units (by closely coupling to the speech envelope) and that higher frequency activity may be critical to decoding the input (Giraud & Poeppel, 2012).

Another type of operation that is ubiquitous in speech is the mapping from acoustic/auditory/phonetic representations to articulatory/motoric representations. For example, given that comprehension significantly precedes production in development, learners must figure out the mapping from the acoustic representation of the new words they acquire to the motoric representation as they learn to accurately produce them. Or, consider that in order to repeat a nonword (which is by definition not stored in the mental lexicon) such as ‘blarfnd,’ a speaker-listener must execute the online coordinate transformation from acoustic–phonetic to articulatory coordinates. These coordinate transformations or sensorimotor mappings, attributed to dorsal stream structures (see Figure 1), have been studied in the feedforward and feedback cases (Cogan et al., 2013; Hickok, 2012) and provide an important link to other research on the links between perception and action.
Finally, the role of linguistic knowledge in online perception is a topic of fundamental interest and importance. Both imaging and electrophysiology are proving to be effective tools in elucidating representational questions (say about lexical and phonological knowledge and representations) and evaluating such theoretically motivated claims. The evidence accumulated thus far suggests that listeners posit abstract linguistic representations that form the basis for their phonological systems; these representations constrain the processing of the speech signal (Kazanina et al., 2006; Näätänen et al., 1997). Knowledge of language guides the predictions that form the basis of perceptual analysis.

See also: INTRODUCTION TO ANATOMY AND PHYSIOLOGY: Auditory Cortex; INTRODUCTION TO CLINICAL BRAIN MAPPING: Disorders of Audition; Language; INTRODUCTION TO COGNITIVE NEUROSCIENCE: Bilingualism; Music; Speech Production; INTRODUCTION TO SYSTEMS: Early Auditory Processing; Grammar and Syntax; Multisensory Integration and Audiovisual Speech Perception; Speech Sounds.

References


