Speech Perception

Lori L. Holt, Ph.D.
Associate Professor
Department of Psychology & Center for the Neural Basis of Cognition
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA USA

Kaori Idemaru, Ph.D.
Assistant Professor
Department of East Asian Languages and Literatures
University of Oregon
Eugene, OR 97405

Lori L. Holt, PhD
Associate Professor
Department of Psychology & Center for the Neural Basis of Cognition
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
Phone: (412)268-4964
Fax: (412)268-2798
Email: lholt@andrew.cmu.edu

Spoken words exist for mere moments, but from this fleeting acoustic signal we are able to apprehend considerable information. We can decode the linguistic message of the speaker as well as information about her gender, age, region of origin, identity and emotional state. As adult listeners, we are so adept at speech perception that the ability seems trivial. However, the ease with which we perceive speech belies the complexity of the perceptual, cognitive and neural mechanisms involved.

The primary reason that speech perception is so complex is there is no straightforward, one-to-one correspondence between a speech segment (e.g., /d/) and its acoustic qualities. About fifty years ago, researchers presumed that there was a simple one-to-one relationship and based on this hypothesis attempted to build a reading machine for the blind whereby written text was translated into a sound alphabet with sound-by-sound translation. However, even with many hours of training, people could not comprehend the machine's speech. This failure led to the discovery that speech is not a sequence of discrete sounds as text is a string of separate letters (Liberman, 1996). Acoustic elements of a spoken word (e.g., the three speech segments in 'dean', /din/) are not produced discretely. Rather, the acoustic information for speech segments overlaps within the acoustic signal such that at any moment of time, the speech signal is colored by the speech uttered before and after. This means that the sound segment /d/, for example, could exhibit many different acoustic signatures depending on its context.

As adult listeners we effortlessly notice that the spoken words *dean, den, dune, and dawn* begin with the same English consonant /d/. This ease belies the complex perceptual processing at work. Not only do the sound properties of/d/ vary with context (such that the

initial sound in *dean* is distinct from that of *dune*) as noted earlier, they also vary with the dialect, gender, emotional state, and physical stature of the speaker. More importantly, there are multiple sound properties associated with the production of /d/ and these properties persist as sound for mere tens of milliseconds. Adding to the complexity, the particular sound properties used to identify speech depend on the listener's native language and expectations. Invariant perception ("these are all/d/s") in the face of variable acoustic signals, therefore, is a remarkable perceptual accomplishment and it has been one of the central puzzles in the study of speech perception.

Early research in speech perception suggested that general auditory processing might not be sufficient to accomplish these perceptual feats. For example, the code of the auditory system seems more variable than the /d/ percept; whereas the sounds in *dune* and *dean* are perceived as equivalent, they seemed to be encoded differently by the auditory system. This led theorists to postulate that the objects of speech perception are not auditory. Instead, the intended articulatory gestures of the speaker (how the speaker configures and coordinates her lips, jaw, and tongue to articulate /d/, for example), as exemplified by the neuromotor command to the articulators, may provide a less variable perceptual code. By this theoretical account, speech perception relies on a specialized perceptual system, distinct from general auditory processing and linked to speech production. The objects of speech perception are the information sound conveys about articulatory events of a speaker rather than the properties of the sound itself. The hypothesized system for accomplishing this process is a modular, innately specified biological specialization for language, distinct from other forms of auditory

processing. This ‘‘speech is special’’ notion was the basis of the Motor Theory of speech perception (Liberman & Mattingly, 1985).

By this view, in hearing /d/ in *dean* and *dune*, for example, the listener recovers the intended neuromotor commands to the articulators that produce /d/ (e.g., place the tip of the tongue at the roof of the mouth between the upper teeth and hard palate). By the view of Motor Theory, these neuromotor commands (or intended gestures) are invariant and relate to the abstract representation of speech. Thus, whereas the sound /d/ may vary in its acoustic realization, all /d/s are perceived as /d/ by virtue of their correspondence to the intended articulatory gesture. In this way, Motor Theorists have attempted to achieve parity between the sending system (production) and the receiving system (perception).

Another prominent theoretical approach, Direct Realism (Fowler, 1986), shares with Motor Theory the claim that the objects of speech perception are articulatory rather than auditory events. However, unlike Motor Theory, Direct Realism denies that specialized processes are necessary to account for speech perception. Rather, following in the tradition of direct realist theories of vision, it asserts that there is rich information available in the speech signal from which listeners may directly recover information, revealing articulatory gestures directly without mediation by cognitive processes of inference or hypothesis testing and without a specialized module as posited by Motor Theory. This theory is realist in the sense that perceivers are thought to recover the actual physical properties of the articulatory gestures from the acoustic signal.

By this view, invariant structure in the acoustic speech signal allows listeners to recover directly the coordinated vocal tract movements that produced the sound /d/. Furthermore, it is

considered that the gesture for /d/ and the gesture for /i/ in *dean* remain as separate and independent perceptual events albeit their temporal overlap in the sound acoustics. By postulating this, Direct Realism accommodates the challenge that articulatory gestures themselves are actually not invariant, but rather are likewise shaped by context.

What Motor Theory and Direct Realism share is the hypothesis that the objects of speech perception are articulatory gestures, the coordinated actions of the vocal tract; listeners do not perceive speech sounds, *per se*. Rather, each of these theories suggests listeners recover information about the patterns of movements the articulators made to produce speech. In this way, these theories draw a line between speech perception and general auditory perception, suggesting that perceiving speech is an act entirely different from perceiving the honk of a car horn, a sequence or Morse code, or the bark of a dog.

However, the theoretical and empirical motivations for positing that speech perception is separate from general auditory perception have weakened in recent years. As noted above, early studies of auditory processing that seemed to indicate too variable an auditory code to accommodate speech perception. However, as researchers focus on how perception of speech is influenced by general perceptual and cognitive capacities of working memory, attention, neural plasticity across different time intervals, and general processing at peripheral and central levels, research is demonstrating that some challenges of speech perception may be met by general auditory perceptual processing. Speech perception may not be inherently different from other types of auditory processing and that investigation of speech perception may inform theories of general perceptual-cognitive processing and vice versa.

This General Approach to speech perception is distinguished from Motor Theory in that it does not invoke specialized mechanisms or modules. Rather, its working hypothesis is that acoustic speech sounds are perceived with the same mechanisms of auditory perception and cognition that have evolved to handle other classes of complex environmental sounds. Further, the General Approach is differentiated from both Motor Theory and Direct Realism in that it assumes that mapping from signal to meaning is not mediated by the perception of articulatory gestures, but rather involves mapping the complex structure of the acoustic signal to regularities learned through experience with the distributions of the ambient language. This general theoretical perspective embraces general perceptual and cognitive mechanisms, not specific to speech, but neither limited to solely low-level sensory processing and psychophysics. The account is "general" in the sense that it suggests that the broad perceptual/cognitive processing of the central nervous system and, as well, the considerable feedback that higher centers have to lower levels of processing are brought to bear in perceiving spoken language (Holt & Lotto, 2008).

By this view, the challenge of perceiving speech in the face of acoustic variability is addressed by listeners' ability to make use of multiple imperfect acoustic cues to learn complex sound categories, like /d/. Auditory processing is sensitive to statistical regularities in the distributions of acoustic attributes as they covary with sound category distinctions. Through this experience, listeners may learn functional equivalence classes of sounds whereby no single acoustic cue is necessary or sufficient to uniquely identify a speech sound, but by which multiple imperfect cues collaborate to relate the variable acoustics to a functional category (e.g., /d/) in the native language.

REFERENCES

Fowler, C. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, 14, 3-28.

Holt, L. L. & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. Current Directions in Psychological Sciences, 17, 42-46.

Liberman, A. M. (1996). Speech: A Special Code. Cambridge, MA: The MIT Press.

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

KEYWORDS

Speech, perception, auditory, phonetic