Jongman, A. (2005). Speech Perception. In P. Strazny (Ed.),
Encyclopedia of Linguistics. New York: Routledge.

The speech stream is a continuously varying signal that, contrary to the listener's impression of silence between words, does not contain any pauses. The primary goal of research in speech perception is to illuminate the way in which the listener converts this continuous signal into a sequence of discrete, meaningful units. This process draws on perceptual, linguistic, and cognitive factors.

At least three major issues need to be taken into account when considering the mapping between acoustic parameters and linguistic units. The first concerns the segmentation of the speech stream. Speech is a highly efficient means of communication in which multiple layers of information are transmitted in parallel. Compare, for example, the words “sea” and “Sue.” The former consists of an [s] consonant followed by a vowel [i], which is pronounced with unrounded lips, whereas the latter consists of the same initial consonant followed by a rounded vowel [u]. During production of the initial consonant in “Sue,” lip rounding will start in anticipation of the upcoming rounded vowel, a process known as *anticipatory coarticulation.* As a result, the acoustic properties of [s] preceding [u] will be different from those of [s] preceding [i]. Thus, the pronunciation of the initial sound in the above words provides acoustic cues about the initial sound itself and about the immediately following sound at the same time. This parallel transmission of information can lead to a segmentation problem in that it is difficult or impossible to chop up the speech signal in chunks that correspond to only a single speech sound. In other words, it is hard--if not impossible--to tell exactly where the consonant ends and the vowel begins.

A second related issue concerns the lack of linearity. Although the listener perceives the speech signal as a linear sequence of units (the size of which is still under debate), the acoustic cues to these units do not necessarily occur in a corresponding left-to-right order in the speech stream. Consequently, acoustic properties appearing later in the speech stream may carry information that is crucial for the identification of an earlier occurring speech sound.

The third issue concerns variability or the lack of invariance. The properties of the acoustic signal that are thought to elicit perception of speech sounds (the acoustic "cues") are never exactly the same. Sources of variability are, for example, differences in vocal tract size, speaking rate, phonetic context, emphasis, and intonation; these can significantly affect acoustic parameters of the speech signal. For example, the most salient frequencies ("formants") of the vowel in the word *heed* are approximately 300 and 2,300 hertz for an adult male and 500 and 3,100 hertz for a child. However, despite these very different formant frequencies caused by differences in vocal tract size, listeners perceive both utterances as containing the vowel [i]. This illustrates what is known as the invariance problem: acoustic cues to a particular speech sound may not be constant but may instead vary according to the circumstances under which they occur. Thus, listeners have to compensate for such differences, or "normalize the input." One of the primary issues in speech perception, then, is how the listener achieves an invariant percept despite great variability in the acoustic input.

Speech perception experiments usually involve manipulated natural speech or synthetic computer-generated speech. Systematic manipulation of individual attributes of speech enables researchers to determine which acoustic properties are necessary and sufficient cues for a particular percept. In applying this methodology to consonants, early research in the 1950s revealed that English listeners use two primary acoustic cues for determining where in the mouth exactly a consonant is articulated: the frequency of the release burst and the formant transitions from the consonant into the following vowel. When the researchers exposed listeners to consonant=-vowel sequences, keeping the burst frequency constant, the listeners perceived a [p] when an [i] was following, but a [k] when an [ɑ] was following. Thus, the exact same cue can participate in different perceptions. In highlighting that individual acoustic attributes are highly context dependent, these experiments also exposed the invariance problem.

Debate continues concerning the extent to which the perception of speech involves the use of biological mechanisms evolved especially for speech. Evidence that speech sounds are perceived differently from their nonspeech analogs was first presented in the early 1960s. These studies specifically examined the way in which these two types of sounds are identified and discriminated. Most types of stimuli (e.g., musical tones, colors) are much better discriminated than they are identified. The greater the physical difference between two stimuli, the better their discrimination. This was shown for nonspeech sounds as well. However, this was not true of certain speech sounds, most notably stop consonants such as [b] or [d]. Discrimination of these sounds was not any better than their identification. For example, in a typical experiment with synthetic speech, an important formant frequency of the consonant was manipulated. [b] typically has a formant at 1,100 hertz, whereas the equivalent formant of [d] lies at 1,800 Hz. With the help of a speech synthesizer, researchers were able to produce sounds with formants that lie somewhere in between. When listeners were asked to identify such intermediate consonants, it turned out that listeners seem to use a particular frequency as the break-off point: all consonants whose formant exceeded this threshold frequency were identified as [d], and all others were identified as [b]. In another experiment, listeners were presented with a pair of consonant stimuli and asked to tell whether the stimuli were the same or different. When the formants of consonants both exceeded or both undercut the threshold frequency, the listener seemed to answer at random. If one consonant fell into the [d] range and the other into the [b] range, however, listeners were consistently able to categorize them properly, even if the formant frequencies were close together. This pattern of results is known as *categorical perception.*

It was originally thought that categorical perception occurred only with speech sounds and not with nonspeech sounds, and this would suggest that the perception of speech engaged specialized mechanisms. However, later experiments with carefully controlled nonspeech materials have shown patterns of categorical perception as well. In addition, animals including the chinchilla, macaque, and Japanese quail have also been shown to have human-like categorical perception. These data concerning categorical perception do not readily support postulation of perceptual mechanisms that were specially evolved or adapted for speech. Instead, they seem to favor an interpretation based on general auditory mechanisms and psychoacoustic sensitivity.

The finding of categorical perception suggested that listeners did not perceive any differences between stimuli belonging to the same category. This, however, may have been based on the particular response categories that were typically used in identification and discrimination experiments (e.g., /b/ or /d/, “same” or “different,” respectively). The use of more sensitive response measures reveals that even though listeners do assign the same label or fail to distinguish between stimuli of the same category, they are in fact aware of subtle differences. These findings indicate that although categories play an important role in the perception of speech, they are not monolithic but have internal structure to which listeners are sensitive.

The importance of speech categories leads to the question of how they are established. This debate centers around the issue of whether speech categories are innate or result from exposure to the ambient language. Crucial evidence in this debate is typically drawn from perception experiments with infants. Findings from discrimination experiments with infants as young as one month old suggest that they divide a speech continuum in a way very similar to adults, with two clearly defined categories and a sharp boundary at the adult location. Additional research has shown that infants up to approximately six months of age can not only discriminate speech categories from their native language, but also from just about any other language, as well. However, in the second half of the first year of life, infants seem to lose their sensitivity to nonnative distinctions. Presumably, the decrease in sensitivity to contrasts that do not play a role in the native language allows for an increase in attention to other aspects of the speech signal that play a role in word learning, such as sentence structure and intonation. Acquisition of speech categories can thus be understood as the result of the interaction between initial psychoacoustically based sensitivities and an increasing awareness of the structure of the language to be learned.

Because the perception of speech draws on many sources of knowledge, theories of speech perception often account for only a few of its components. Three general classes of models may be distinguished. The motor theory of speech perception deals with acoustic variability by claiming that the listener has specialized neural mechanisms to convert the speech signal into invariant representations of articulatory gestures. These articulatory gestures are the object of speech perception; i.e., this theory assumes that the listener attempts to faithfully reconstruct how the perceived speech sounds were articulated by the speaker. The theory of acoustic invariance claims that invariant acoustic properties do reside in the speech signal. By using specialized neural mechanisms, the listener directly extracts these invariants from the speech signal and maps them onto phonetic features. Finally, pattern recognition models claim that speech perception is much like statistical pattern classification. No specialized mechanisms are required. Instead, the unit of recognition and the structure of categories (e.g., based on prototypes or exemplars) is determined by the nature of the speech signal and general properties of the mammalian auditory system.

There is a growing recognition that a detailed analysis of the speech signal alone will not be sufficient to obtain a genuine understanding of the way in which speech is perceived. Consequently, researchers have started to incorporate findings from additional areas. One such area concerns the way in which the speech signal is transformed by the auditory system. Although the measurements used in phonetic analysis typically represent frequency along a linear scale, it is known that the auditory system warps the signal such that its ultimate representation is more nearly logarithmic in nature. A thorough understanding of these transformations at the auditory periphery and higher levels along the auditory pathway may well have significant implications for the current view of acoustic cues and their variability.

A second area concerns the relation between the speech signal and higher levels of organization of the grammar. The first few decades of research on speech perception emphasized the cataloguing of acoustic cues, that is, the acoustic information that listeners need to extract from the speech signal to recognize individual speech sounds. Since the 1980s, however, increased interest in the way words are recognized has led to research on what is often referred to as spoken or auditory word recognition. A central issue for this research area is if and to what extent "higher-level" linguistic and cognitive information that is not present in the speech signal contributes to word recognition. Research has shown that when presented with a sequence of speech sounds containing an ambiguous initial consonant, listeners will classify that consonant such that the entire string will result in an existing word instead of a nonword. For example, listeners will report hearing "beef" rather than "peef" when presented with the string "eef" preceded by a sound that is ambiguous between /b/ and /p/. Conversely, listeners will classify the same ambiguous initial consonant as /p/ rather than /b/ when it is followed by "eace." Findings such as these are often considered as evidence that lexical information (knowledge about what constitutes a word) affects the listener's interpretation of acoustic=-phonetic information. The extent to which there is feedback from higher levels of linguistic representation, such as the lexicon, is still very much under debate.

Allard Jongman

Further Reading

Gernsbacher, Morton Ann, editor, *Handbook of Psycholinguistics*. San Diego, California: Academic Press, 1994

Goodman, Judith C., and Howard C. Nusbaum, editors, *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words.* Cambridge, Massachusetts: MIT Press, 1994

Hardcastle, William, and John Laver, editors, *The Handbook of Phonetic Sciences.* Oxford and Cambridge, Massachusetts: Blackwell, 1997

Harnad, Steven, editor, *Categorical Perception: The Groundwork of Cognition*. Cambridge: Cambridge University Press, 1987

Jusczyk, Peter, *The Discovery of Spoken Language.* Cambridge, Massachusetts: MIT Press, 1997

Kuhl, Patricia, “On Babies, Birds, Modules, and Mechanisms: A Comparative Approach to the Acquisition of Vocal Communication,” in *The Comparative Psychology of Audition: Perceiving Complex Sounds,* edited by Robert J. Dooling and Stewart H. Hulse, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1989

Liberman, Alvin M. *Speech: A Special Code*. Cambridge, Massachusetts: MIT Press, 1996

Marslen-Wilson, William, editor, *Lexical Representation and Process.* Cambridge, Massachusetts: MIT Press, 1989

Moore, Brian C. J., editor, *Hearing*. San Diego, California: Academic Press, 1995

Miller, Joanne L., Raymond D. Kent, Bishnu S. Atal, editors, *Papers in Speech Communication: Speech Perception*. Woodbury, New York: Acoustical Society of America, 1991

Nygaard, Lynne C., and David B. Pisoni, “Speech Perception: New Directions in Research and Theory,” in *Speech, Language, and Communication,* edited by Joanne L. Miller and Peter D. Eimas, San Diego, California: Academic Press, 1995

Schouten, Martin E.H., editor, *The Psychophysics of Speech Perception.* Dordrecht: Kluwer, 1987

Warren, Richard M. *Auditory Perception: A New Analysis and Synthesis*. Cambridge: Cambridge University Press, 1999