Speech Perception Within an Auditory Cognitive Science Framework

Lori L. Holt[1] **and Andrew J. Lotto**[2]

[1]*Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, and*
[2]*Department of Speech, Language, and Hearing Sciences, The University of Arizona*

ABSTRACT—***The complexities of the acoustic speech signal pose many significant challenges for listeners. Although perceiving speech begins with auditory processing, investigation of speech perception has progressed mostly independently of study of the auditory system. Nevertheless, a growing body of evidence demonstrates that cross-fertilization between the two areas of research can be productive. We briefly describe research bridging the study of general auditory processing and speech perception, showing that the latter is constrained and influenced by operating characteristics of the auditory system and that our understanding of the processes involved in speech perception is enhanced by study within a more general framework. The disconnect between the two areas of research has stunted the development of a truly interdisciplinary science, but there is an opportunity for great strides in understanding with the development of an integrated field of auditory cognitive science.***

KEYWORDS—***speech perception; auditory perception; auditory cognitive science***

The ease with which a listener perceives speech in his or her native language belies the complexity of the task. A spoken word exists as a fleeting fluctuation of air molecules for a mere fraction of a second, but listeners are usually able to extract the intended message. The seemingly trivial ability to determine that the spoken words *dean*, *den*, *dune*, and *dawn* begin with the same English consonant, /d/, is actually a remarkable accomplishment. There are a number of physical acoustic characteristics associated with the production of /d/, and they unfold quickly across just tens of milliseconds. These acoustic correlates vary with the following vowel (such that the initial sound in *dean* is distinct from that of *dune*) and the preceding word context, as well as with the dialect, gender, emotional state, and physical stature of the speaker. To make matters more challenging, the particular acoustic correlates utilized by a listener depend on that listener's native language and expectations. Invariant perception ("these are all /d/s") in the face of variable acoustic signals has been one of the central puzzles in the study of speech perception for more than 50 years.

Presumably, understanding how the auditory system processes complex sounds can be informative about how speech perception is accomplished; and, likewise, investigation of speech perception can provide clues to how the auditory system functions. Unfortunately, such a symbiosis has yet to be realized fully because there has been segregation of inquiry into speech perception and general auditory perception.

Early in the development of the field of speech perception, researchers noted that the auditory system encodes the /d/ in *dean* differently than the /d/ in *dune*. The fact that listeners' percepts were more invariant than would be expected from the variable auditory code suggested that auditory processing did not sufficiently constrain speech perception. Theorists reconciled the invariant speech perception with the variable acoustic signatures and auditory encoding by proposing that speech perception relies on a specialized perceptual system distinct from general auditory processing. This hypothesized system was modular, in the sense that its operations were hypothesized to be dedicated to perceiving sounds produced by the human voice and to be impenetrable to influence from more general processing. This "speech is special" notion was codified in the Motor Theory of speech perception (Liberman & Mattingly, 1985).

The theoretical and empirical motivations for separating the study of speech perception from general auditory science have weakened in recent years. More and more, researchers focus on how perception of speech is influenced by working memory, attention, neural plasticity across different time intervals, and

Address correspondence to Lori L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213; e-mail: lholt@andrew.cmu.edu.

general processing at peripheral and central levels. Research in both speech and general audition indicates the promise of cross-fertilization. A new conceptualization is forming that (a) speech perception is not inherently different from other types of auditory processing, and that (b) investigation of speech perception informs theories of general perceptual-cognitive processing and vice versa.

The time is right to develop more fully an interdisciplinary approach to auditory perception and cognition, with perception of speech as a central focus. Here we present examples of how general auditory and speech research may mutually enhance one another. We conclude with suggestions of the kinds of research questions that become viable within a comprehensive auditory-cognitive framework.

PHONETIC CONTEXT EFFECTS

Speech sounds like the paradigmatic /d/ possess a complex acoustic structure. When people articulate, their vocal-tract configurations emphasize some frequencies and attenuate others. As a result, one can partly characterize speech by its patterns of high-energy peaks, or *formants*, across frequencies. These patterns change across time and serve as information about speech-sound identity. The consonant of *da*, for example, may be distinguished from that of *ga* by examining the formant patterns shown in Figure 1. Note that the patterns are very similar except that, in the higher frequency range, one of the formants (the third formant, F3) begins at a higher frequency for *da* and at a lower frequency for *ga*. This pattern is relatively consistent across speakers and dialects. If F3 were set at a frequency between these two extremes, perception would vacillate between *da* and *ga* across presentations. To this point, recognizing /d/ versus /g/ appears a simple issue of auditory pattern recognition.

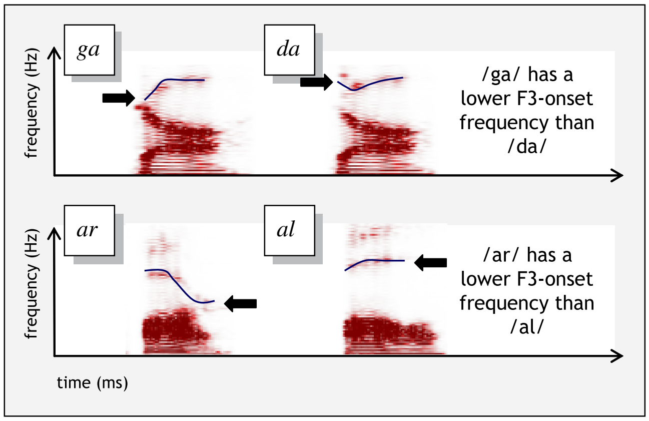

Fig. 1. Spectrograms showing the formant patterns for the sounds *ga*, *da*, *ar*, and *al*. Formants are patterns of high-energy peaks; sound frequency is plotted against time from sound onset and continuing through offset; darker colors illustrate higher-amplitude sound. Note that patterns for *ga* and *da* are similar except that, in the higher-frequency range, one of the formants—the third formant or F3 (tracked with blue lines)—starts relatively low for *ga* and relatively higher for *da*. The patterns for *ar* and *al* are also similar but *ar* has a lower frequency F3 offset than *al*.

However, F3 frequency is not only a function of whether a speaker is producing a /d/ or /g/; it is also influenced by the speech that precedes (and follows) the consonant. Fluent speech demands rapid articulation, but the human vocal tract cannot move instantaneously. Thus, the actual placement of the tongue (or jaw, etc.) in producing a consonant is influenced by where the tongue was before and where it is going next. This context-dependent production is called *coarticulation*, and it influences the acoustics (including F3) of the resulting sound. Coarticulation presents a major problem for conceptualizing speech perception as auditory pattern recognition; the acoustic correlates of /d/, for example, shift radically as a function of surrounding sounds.

How do listeners deal with coarticulation? Mann (1986) demonstrated that perception of *da* and *ga* shifts as a function of preceding context. If a consonant with an ambiguous F3 frequency is preceded by a word ending in /al/, such as *fall*, listeners perceive the ambiguous target as *ga*. If the same consonant is preceded by /ar/, as in *far*, listeners hear *da*. This context-sensitive speech perception is the complement to effects of coarticulation. Following /al/, speech production is more /da/-like (with the tongue shifted toward the front of mouth), but *perception* following /al/ is more /ga/-like. Perception thus appears to compensate for coarticulation, suggesting the possibility that listeners use knowledge specific to speech production to disentangle effects of coarticulation and recover the intended production. By this classic explanation, specialized modular speech processors are hypothesized to compensate for coarticulation.

The tight link between coarticulation and perceptual compensation appears to be strong evidence for perceptual processes specific to speech. However, the same pattern of perception can result from general processes. When tones mimicking *al* and *ar* F3 frequencies precede syllables that vary perceptually from *da* to *ga*, listeners' speech perception shifts (Lotto & Kluender, 1998). These tones sound nothing like speech and presumably do not engage speech-specific processes. Yet they shift speech perception in the same manner as the syllables they model. Further, birds trained to peck to /da/ versus /ga/ shift their pecking behavior when the targets are preceded by /al/ or /ar/, despite the fact that birds are quite unlikely to have speech-specific processing modules (Lotto, Kluender, & Holt, 1997). Context-dependent speech perception is not specific to speech contexts or to human listeners.

What accounts for this generality? Recall from Figure 1 that /da/ has a high-frequency F3 onset whereas /ga/ has a low-frequency F3 onset. Figure 1 also illustrates that, similar to the acoustic correlates for /da/ versus /ga/, /al/ has a high F3 offset and /ar/ has a low F3 offset. One can describe the perceptual context dependence in relative terms: After a high F3 (or high tone), an ambiguous F3 is perceived as relatively lower in frequency (more *ga*-like). The auditory system appears to represent acoustic signals not in terms of absolute values, but relative to sounds that precede (and follow) them (Wade & Holt, 2005). This

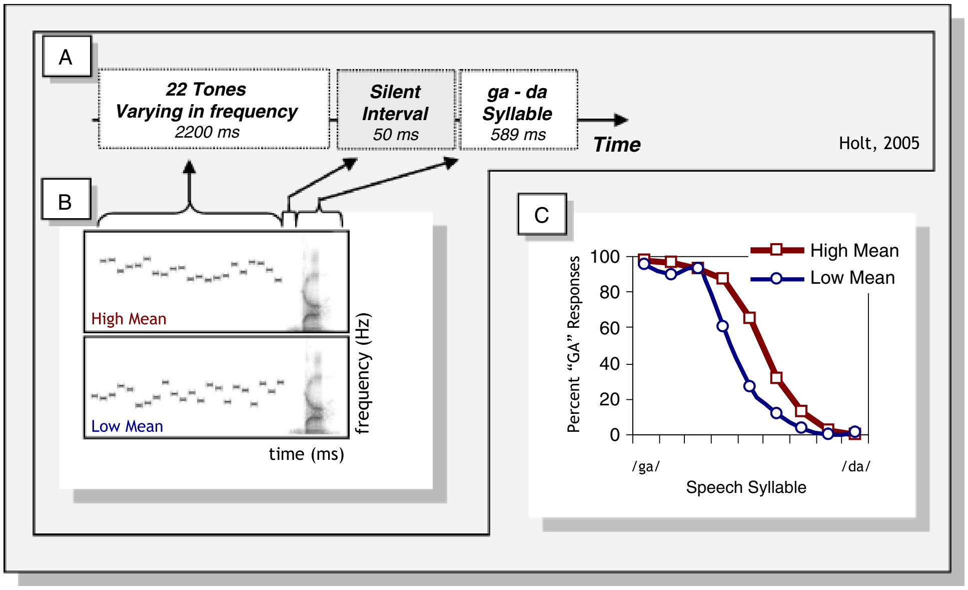

Fig. 2. Experiment showing the sensitivity of speech categorization to nonspeech context (Holt, 2005). The schematic shown in (A) illustrates the stimuli used: 22 tones varying in frequency formed a sequence of sounds that preceded a syllable drawn from a series that varied perceptually from *ga* to *da*. The tones had either a high-frequency distribution or a low-frequency distribution, as shown by two samples plotted on time-by-frequency axes in (B). When listeners identified the speech syllables in these tone-sequence contexts, the distribution of tone frequencies had an effect on speech perception (C). Listeners more often identified the consonants as *ga* when high-frequency sequences of tones preceded the syllables. The same syllables were more often reported to be *da* when low-frequency sequences preceded them.

context sensitivity is a consequence of the general operating characteristics of human (and bird) auditory systems. Some of the perceptual complexity introduced by coarticulation may be accommodated by very general auditory processing.

The implications of context-sensitive audition extend beyond speech perception. If perception of complex sounds depends on context, then studying the auditory system using isolated stimuli will not provide comprehensive understanding. We know little about how neighboring sounds influence auditory encoding and representation. Our findings from studies of speech perception show the need for research on context effects in auditory perception and suggest tools for such investigation.

Our research is already demonstrating the fruitfulness of integrating investigation of speech and general auditory processing. In examining the extent of context-sensitive audition, Holt (2005) presented listeners with targets varying perceptually from *da* to *ga*, preceded by a sequence of 22 tones of different frequencies (Fig. 2A). For some trials, the distribution of tones forming the sequence had a high average frequency; for others, it had a low average frequency (Fig. 2B). Listeners' perception of syllables following these tone sequences was context sensitive; following the high-frequency tone distribution, syllables were identified more often as *ga*, the lower-F3 alternative (Fig. 2C). A series of follow-up experiments suggested that the auditory system maintains a running estimate of context and encodes incoming sound targets relative to that estimate (Holt, 2006). This running estimate lasts more than a second and is not disrupted by intervening neutral-frequency tones, noise, or silence. These results suggest that a memory buffer across seconds, the calculation of statistics from acoustic events, and representation of the target relative to a statistical standard may be integral components in perception of speech and other complex sounds. The investigation of these components extends outside the realm of traditional research in speech or audition and requires a more interdisciplinary cognitive-science approach to reveal their bases.

SPEECH AND EXPERIENCE

Audition is dependent on short-term experience, as the example above illustrates; but longer-term experience with regularities of speech is also important. Speech sounds are grouped by functional significance within a language; for instance, /l/ and /r/ are distinct in English but not in Japanese. Experience with these regularities tunes perception such that identical acoustic signals may be perceived differently by listeners with different language experience. These changes are thought to reflect functional grouping of speech sounds as categories.

Although experience is clearly important in developing speech categories, it has proven difficult to determine underlying learning mechanisms, because adults (and even infants) already possess a great deal of speech experience, precluding precise characterization or experimental control of experience. The enterprise of investigating the underlying learning mechanisms important for speech has benefited from study within a broader cognitive-science framework.

Nonhuman animals, for example, have been found to respond to the statistical regularities of speech experience just as human adults and infants do (e.g. Holt, Lotto, & Kluender, 2001; Hauser, Newport, & Aslin, 2001). Likewise, training human adults to categorize complex nonspeech acoustic stimuli that model some of the complexity of speech stimuli has demonstrated that there are important perceptual (Holt, Lotto, & Diehl, 2004) and cognitive (Holt & Lotto, 2006) constraints on auditory learning that extend to speech categorization. Considering speech perception from a general cognitive-science perspective, it also becomes possible to integrate findings from visual categorization into theories of speech. There may be many parallels, for example, between speech perception and development of expertise for faces. Reuniting investigation of speech and general perceptual/cognitive processing provides converging methods to investigate auditory learning, plasticity, and development of expertise.

AN AUDITORY COGNITIVE SCIENCE

In many ways, the disconnect between speech and auditory perception has stunted development of a truly interdisciplinary auditory cognitive science. The examples above highlight reciprocal benefits of the unconstrained study of speech perception within a general auditory-cognitive framework. In attempting to understand how auditory processing influences the perception of speech, we also may use speech to gain insight into auditory cognition.

Consideration of perceptual compensation for coarticulation in an auditory cognitive-science framework, for example, makes new general auditory research questions viable. Finding that a statistically defined sequence of tones influences listeners' speech categorization raises as-yet-unanswered questions: To what kinds of statistical regularity is auditory processing sensitive? What is the time course of the memory buffer and are these online "running estimates" related to working memory? Do mid-level perceptual processes like perceptual grouping influence the units across which regularities are computed? How do these effects interact with higher-level linguistic processing?

Within the framework of an auditory cognitive science, it would be a mistake to presume that speech perception is necessarily encapsulated or isolated from other perceptual and cognitive processes. There is no a priori reason to propose that speech perception does not receive input from senses other than audition (in fact there is substantial evidence to the contrary) or that speech perception is a process that takes basic sensory input and outputs a string of phonemes. It is either an explicit or implicit presumption of most models of spoken-language perception that linguistic processing begins with discrete phoneme representations, isolated from the messy acoustics involved in their characterization. However, research has undermined this view, as effects of acoustic variation are evident in word recognition (e.g., Hawkins, 2003) and expectations from word and semantic representations feed back to influence speech-sound perception (McClelland, Mirman, & Holt, 2006). Indeed, it is difficult to determine where auditory perception ends and cognition begins; the perception/cognition boundary may be artificial, and assuming its existence may be counterproductive to progress in understanding.

It is worth noting that acceptance of a general cognitive framework for speech perception does not demand dismissal of all processes purported to be specialized for speech signals. It may be the case that evolved mechanisms or learned processes enacted only on signals that resemble speech play a role. Nor does this perspective negate the importance of speech production in understanding speech communication. Perception and action are tightly coupled in human behavior, but this relationship does not dictate that motor representations are primary in speech perception, nor does it necessitate specialized processing. We are proposing that these questions be tested within a general cognitive-perceptual framework, with skeptical conservatism in interpretation of phenomena before inferring specialized mechanisms or modules.

The study of speech has long been relegated to the periphery of cognitive science, as a "special" perceptual system that could tell us little about general issues of human behavior. Given continued development of new tools for the synthesis and manipulation of complex sounds, along with innovations in the ability to examine human neural processing, the prospects for exploiting speech in development of an auditory cognitive science are very encouraging. Marshalling the skills and expertise of researchers across disciplines, and integrating findings from other auditory research domains like music and auditory scene analysis, being able to understand the seemingly simple act of recognizing a /d/ amidst the variability of the speech signal has much to offer cognitive science.

Recommended Reading

Diehl, R.L., Lotto, A.J., & Holt, L.L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179. An accessible overview of important phenomena and theories of speech perception.

Liberman, A.M. (1996). *Speech: A special code*. Cambridge, MA: The MIT Press. A comprehensive review of research and theorizing relevant to the motor theory of speech perception.

McAdams, S., & Bigand, E. (Eds.) (2001). *Thinking in sound: The cognitive psychology of human audition*. Oxford, UK: Oxford University Press. An edited volume surveying the cognitive psychology of audition.

Acknowledgments—Supported by National Institutes of Health Grant R01DC004674-04A2 to the coauthors.

REFERENCES

Hauser, M.D., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*, B53–B64.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405.

Holt, L.L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science*, *16*, 305–312.

Holt, L.L. (2006). The mean matters: Effects of statistically-defined non-speech spectral distributions on speech categorization. *Journal of the Acoustical Society of America*, *120*, 2801–2817.

Holt, L.L., & Lotto, A.J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, *119*, 3059–3071.

Holt, L.L., Lotto, A.J., & Diehl, R.L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *Journal of the Acoustical Society of Americav*, *116*, 1763–1773.

Holt, L.L., Lotto, A.J., & Kluender, K.R. (2001). Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement? *Journal of the Acoustical Society of America*, *109*, 764–774.

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

Lotto, A.J., & Kluender, K.R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*, 602–619.

Lotto, A.J., Kluender, K.R., & Holt, L.L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America*, *102*, 1134–1140.

Mann, V.A. (1986). Distinguishing universal language-specific factors in speech perception: Evidence from Japanese listeners' perception of /1/ and /r/. *Cognition*, *24*, 169–196.

McClelland, J.L., Mirman, D., & Holt, L.L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Science*, *10*, 363–369.

Wade, T., & Holt, L.L. (2005). Effects of later-occurring non-linguistic sounds on speech categorization. *Journal of the Acoustical Society of America*, *118*, 1701–1710.