EUROPEAN STUDIES IN PHONETICS AND SPEECH COMMUNICATION

Edited by
Gerrit Bloothooft
Valerie Hazan
Dieter Huber
Joaquim Llisterri

1995

Directions in Speech Perception Research

Doug H. Whalen
Haskins Laboratories, New Haven, CT, USA

Language is a biological system and as such brings a complex evolutionary history into the present. Humans are predisposed to produce and perceive speech, and they acquire by two years of age a system that we have as yet been unable to duplicate in machines. The evidence points to a neurological specialization for speech. Such a specialisation makes humans typical members of the biological world: Behaviors that become communicative acts in animals tend to come under separate neurological control (Wilson, 1975), and it is likely that every species has a specialization for communication with its own kind.

In spite of the evidence for such a biological basis, the most common views of speech assume that it is a cognitive process of matching sensory stimulation to categories formed by habitual association. Phonological segments, on this account, are not biological categories but simply the prototypes built up from the constant exposure to speech that humans encounter. The powerful rule-generating system that language is assumed to be then allows these prototypes to be associated with words and then on to the rest of language.

This division between the biologically based theories and the cognitively based ones has existed for more than a quarter century and looks (despite the premature obituary in Nooteboom's contribution to this volume) as though it will continue for some time. Most of the topics studied, though, are of interest to both kinds of theories, which allows the different perspectives to illuminate the same kinds of results. There are many issues currently under investigation, from perception in infants through changes in old age, from speech coding in cochlear implants to effects of native language on perceiving foreign contrasts, etc., but I will highlight only three: the parsing problem (or, what are we really hearing?); duplex perception (or, how can we hear two things at once?); and brain imaging (or, where is speech perception really taking place?)

The parsing problem

The parsing problem is this: How is it that we get seemingly invariant phonological categories out of a continuous speech stream, that is, how can we "parse" speech? This has been an enduring problem since it was first discovered instrumentally that there is no one-to-one correspondence between phonemes and anything in the speech signal (Potter, Kopp, and Green, 1947; Joos, 1948). This influence of one segment on another has come to be known as coarticulation, and the possible explanations of how humans deal with it have been numerous. Indeed, even though coarticulation is described as an interaction among segments, it has not been established that the segment is the primary unit in perception (or, in some theories, even a necessary one). While parsing and the determination of units are interrelated, it is possible to study them separately.

One of the most appealing attempts to deal with coarticulation has been the auditorily based feature detector (Stevens and Blumstein, 1978; Blumstein, Isaacs, and Mertus, 1982). In theory, these detectors operate on simple acoustic properties that are responded to directly by the auditory system, giving as their output the distinctive features that are needed for phonology. However, the accuracy of the proposed detectors has never approached that of human listeners, and human listeners do not respond appropriately to stimuli designed to fit precisely into the parameters of the detectors (for a review: see Whalen, 1991). In addition, listeners are sensitive to phonetic information even when the feature detectors would seem to ignore such information (Martin and Bunnell, 1982; Whalen, 1984). An alternative is called for.

Another approach holds that listeners do not perceive the acoustic features that make up the immediate stimulus, but rather the phonetic gestures that went into the making of that signal (Fowler, 1980; Liberman, and Mattingly, 1985). In such theories, the changes induced by coarticulation are not random events which make perception hard, but rather information about the segmental context. In an elegant set of experiments, Fowler (1984) demonstrated that listeners do indeed "parse" the signal so that the effects of the segment inducing the coarticulatory changes are attributed to that segment: A coarticulated segment was matched to an acoustically different variant (from a different coarticulatory environment) more often than to the acoustically identical version out of context.

Presumably, a complete description of the parsing process will also tell us what the units of speech are - currently, we have evidence that any number of units play a role in speech perception, with no clear indication that any one is "the" unit of perception. We do not know whether some portion of the speech signal is taken as a beginning point, with other sounds being interpreted in relation to it, or if perhaps all of the possible parsings of a signal arise simultaneously. We can expect to see more progress in this area, probably with the development of new paradigms and techniques.

Duplex perception

For normal conditions, we actually receive two auditory signals rather than one. Each ear receives a slightly different signal, and those differences can be used to let the listeners know where a sound came from ("auditory localisation"). If the signals appear to be two different acoustic events, though, then it just sounds like one signal at one ear and another at the other. Duplex perception is a case where a signal at one ear is used to form percepts at both ears. First described by Rand (1974), the technique involves taking a small part of a syllable (the "extract") and playing it to one ear, while the rest of the syllable (the "base") goes to the other ear. The base by itself is an ambiguous syllable. When it is played to one ear simultaneously with the excerpt at the other, the base results, not surprisingly, in the percept of a syllable at its ear. At the other ear, the excerpt is heard as a nonspeech "chirp". But, oddly, the syllable is no longer ambiguous, but instead has its consonant determined by the information in the excerpt. The excerpt evokes a duplex percept, forming part of two percepts at once.

This bizarre manipulation has told us a lot about speech. Discrimination based on the speech is quite different from that based on the nonspeech, even though the physical signal is, of course, identical (Mann and Liberman, 1983). Speech is an extremely powerful form of organization, taking precedence over the nonspeech system (Whalen and Liberman, 1987), and robust in the face of many different competing signals (Ciocca and Bregman, 1989; Whalen and Liberman, in press). (Vowels seem to be less able to withstand competition; see Darwin and Sutherland, 1984). Even more amazing is the behaviour of 3- and 4-month-old infants: With duplex stimuli, they discriminate the speech quite well, but the nonspeech not at all (Eimas and Miller, 1992).

Duplex perception exploits an unnatural stimulus to illuminate the processes of speech perception and has been one of the most difficult challenges to theories that would rely on auditory coherence alone to organise our perceptual world. Further work will probably include making the nonspeech half of the duplex stimuli form part of a more natural sound, so that the speech part of the duplex percept does not have the advantage of being produced by a real system while the excerpt does not (Fowler and Rosenblum, 1990).

Brain imaging

No matter what theory of speech perception one has, it is clear that the brain performs the actual work. Recent advances in our ability to see the brain in action will provide a wealth of material for future theorists to work with. (The use of these

methods in word recognition is discussed in Cutler's contribution to this volume.) The most difficult part of working with these new techniques is being certain that we know what mental processes are active in the experimental situation. The careful selection of comparison tasks is important in this regard.

Positron emission tomography (PET) reveals activation in particular brain regions by tracking radioactive isotopes in the blood. In one study, Zatorre, Evans, Meyer, and Gjedde (1992) found that phonetic discrimination led to increased activity in Broca's area in the left hemisphere, while a pitch judgment on the same syllabic stimuli activated the right prefrontal cortex. Other PET studies have been performed by Demonet and colleagues (Demonet, Chollet, Ramsey, Cardebat, Nespoulous, Wise, et al., 1992: Demonet, Price, Wise, and Frackowiak, 1994). To some extent, these results simply confirm in the normal subject what we have already inferred from the study of aphasia. Further refinements should allow us an even better window into the working mind.

Magnetic Resonance Imaging (MRI) is another imaging technique that is becoming more useful in studying speech perception. MRI has the advantage that no radio-isotopes are used, and so the length of the sessions can be longer. One recent study (Shaywitz, Shaywitz, Pugh, Constable, Skudlarski, Fulbright, et al., 1995) found that males and females performed the same linguistic task using different parts of their brains: Males primarily used the left inferior frontal gyrus to determine whether two words rhymed, while females had more diffuse and bilateral activation.

Prospect

We are far from knowing everything there is to know about speech perception. The coming years hold the promise of providing answers to some of our long-standing questions and allowing us to ask others that were unaddressable before. The results will change our use of computers, our conception of the mind, and our assumptions about evolution. What more could you want?

Acknowledgments

The writing of this paper was supported by NIH grant HD-01994 to Haskins Laboratories. Ken Pugh, Margaret Hall Dunn, and Ocke-Schwen Bohn provided helpful comments.

References

Blumstein, S.E., Isaacs, E., and Mertus, J. (1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America*, 72, 43- 50.

Ciocca, V. and Bregman, A.S. (1989). The effects of auditory streaming on duplex perception. *Perception and Psychophysics*, 46, 39-48.

Darwin, C.J. and Sutherland, N.S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology*, 36A, 193-208.

Demonet, J.F., Chollet, F., Ramsey, S., Cardebat, D., Nespoulous, J.L., Wise, R., Rascol, A., and Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain*, 115, 1753-1768.

Demonet, J F., Price, C., Wise, R., and Frackowiak, R. (1994). A PET study of cognitive strategies in normal subjects during language tasks: Influence of phonetic ambiguity and sequence processing on phoneme monitoring. *Brain*, 117, 671-682.

Eimas, P.D. and Miller, J.D. (1992). Organization in the perception of speech by young infants. *Psychological Science*, 3, 340-345.

Fowler, C.A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics*, 8, 113-133.

Fowler, C.A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, 36, 359-368.

Fowler, C.A. and Rosenblum, L.D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 742-754.

Joos, M. (1948). *Acoustic phonetics*. Baltimore, MD: Linguistic Society of America.

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

Mann, V.A. and Liberman, A.M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, 14, 211-235.

Martin, J.G. and Bunnell, H.T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 473-488.

Potter, R.K., Kopp, G.A., and Green, H.G. (1947). *Visible speech*. New York: Van Nostrand.

Rand, T.C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, 55, 678-680.

Shaywitz, B.A., Shaywitz, S.E., Pugh, K.R., Constable, R.T., Skudlarski, P., Fulbright, R.K., Bronen, R.A., Fletcher, J.M., Shankweiler, D.P., Katz, L., and Gore, J.C. (1995). Sex differences in the functional organization of the brain for language. *Nature*, 373, 607-609.

Stevens, K.N. and Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64, 1358-1368.

Whalen, D.H., and Liberman, A.M. (1987). Speech perception takes precedence over nonspeech. *Science*, 237, 169-171.

Whalen, D.H. (1984). Subcategorical mismatches slow phonetic judgments. *Perception and Psychophysics*, 35, 49-64.

Whalen, D.H. (1991). Perception of the English /s/ -/S/ distinction relies on fricative noises and transitions, not on brief spectral slices. *Journal of the Acoustical Society of America*, 90, 1776-1785.

Whalen, D.H. and Liberman, A.M. (in press). Independence of auditory source assignment and the speech module. *Haskins Laboratories Status Report*.

Wilson, E.O. (1975). *Sociobiology*. Cambridge, MA: Harvard University Press.

Zatorre, R.J., Evans, A.C., Meyer, E., and Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, 256, 846-849.