1016

Introduction to papers on speech recognition and perception from an articulatory point of view

Richard S. McGowan and Alice Faber
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

(Received 11 August 1995; accepted for publication 20 September 1995)

The following group of papers resulted from a special session entitled *Speech Recognition and Perception from an Articulatory Point of View* that was held during the spring 1994 meeting of the Acoustical Society of America in Cambridge, Massachusetts. Organization of the session began when Richard McGowan, Terry Nearey, and Juergen Schroeter invited speakers to give papers and critiques on the role of articulation in human perception and machine recognition. Presentations were invited from three speakers or groups of speakers who were representative of the three speech areas in the Society: production, perception, and processing. One talk was given by Björn Lindblom, another was given by John Ohala, and the third by Rick Rose, Juergen Schroeter, Mohan Sondhi, and Oded Ghitza. The invited critiquers for the Lindblom paper were Ken Stevens, Robert Remez, and Bishnu Atal; for the Rose *et al.* paper, Roger Moore, Joe Perkell, and Terry Nearey; and for the Ohala paper, Mary Beckman, Douglas O'Shaughnessy, and Carol Fowler. These people created a very interesting and provocative session. Robert Fox, associate editor of the speech perception section of the *Journal*, initiated publication of and edited the papers that appear here. Written versions of the invited papers are included here, as well as the critiques by Stevens, Remez, Nearey, Moore, and O'Shaughnessy. Because it was felt that Carol Fowler's view of the importance of articulation in speech perception was not sufficiently aired in a short critique, she was asked to present her views in a longer paper.

PACS numbers: 43.71.An

The four papers in this section take partially overlapping approaches to the role of recovery of articulation in speech perception by humans and machines. Lindblom stresses that the goal of ordinary speech perception is comprehension of a message, not recovery of a sequence of phonological units, however defined. Just as speakers are guided in their speech by their expectations about their listeners' knowledge and experience, so too listeners bring their knowledge of the world and their linguistic experience to bear on understanding spoken language.

Ohala, in typically lively fashion, discusses constraints on phonological inventories in a variety of languages that arguably are based on the acoustic/auditory properties of the sounds in question. While acknowledging that humans *are* capable of recovering articulatory information from acoustic signals (otherwise, nobody would ever learn to speak!), Ohala argues from these constraints that ordinary speech perception involves recovery of precisely those acoustic characteristics that these observed constraints are stated in terms of.

Fowler, after reviewing crucial differences between the motor theory of speech perception and her own direct realist theory, two theories of speech perception that agree in hypothesizing that speech perception involves recovery of articulatory units, situates speech perception in a more general theory of perception. She argues that theories of speech perception must be compatible with theories of speech production. She further surveys five sets of experimental findings which can be explained in terms of either articulatory primacy in perception or some alternative accounts; because the nonarticulatory explanations for diverse findings—the McGurk–MacDonald effect, the invariant percept /d/ arising from acoustically disparate /di/ and /du/, and trading relations—do not converge, the fact that all can be accounted for with a single, articulation-based theory of speech perception should lead to acceptance of that theory.

Rose, Schroeter, and Sondhi address the pragmatic aspects of using articulatory knowledge in automatic speech recognition (ASR), and not the issue of how humans perceive speech (see Nelson and Bourlard, 1995 for a recent perspective on speech recognition). They appear to favor the use of articulatory knowledge in speech recognition, but point to two obstacles: "...the difficult problem of acoustic-to-articulatory mapping, and the lack of efficient formalisms for combining articulatory knowledge with empirical observation... ." Two alternative directions are indicated for using articulatory knowledge: recovery of explicit articulatory representation from the speech acoustics and followed by recovery of the linguistic content from that representation, and incorporation of knowledge of articulatory behavior into automatic speech recognition model structures. This latter method would be a way of making statistical modeling more knowledge driven and less data driven.

One theme that recurs in the four target papers and in many of the commentaries is sound change. Cataloging and categorizing sound changes have been central to historical linguistics for more than a century. Attempts to explain *why* sound change occurs, both in the particular case (e.g., why did the Middle English vowel *u: diphthongize to /au/, giving the pronunciation /haus/ for *house* instead of older /hu:s/?) and the general case (why does pronunciation change at all?) have likewise attracted much scholarly interest (e.g., Hock, 1986; Kiparsky, 1988; Faber, 1992;

 0001-4966/96/99(3)/1680/3/$6.00

McMahon, 1994). As Lindblom notes, accounts of sound change generally distinguish between two types: listener-based sound change and speaker-based sound change. The two types share several characteristics. To oversimplify only slightly, speaker-based sound change tends to apply to all sounds of a particular class, and to involve articulatory simplifications. Nasalization of a vowel before a nasal consonant is a prototypical sound change of this sort. In contrast, listener-based sound changes tend to be sporadic and to involve listener inference of an articulation other than that actually made by the speaker. Ohala's example of spontaneous vowel nasalization in Breton and in Hindi is an example of this sort of sound change. Unlike the nasalization of /æ/ in English *can't*, the Hindi and Breton nasalization is not triggered by a following nasal consonant.

Within the context of the present grouping of papers, there is little disagreement about the facts of sound change; however, the present authors disagree in how much weight to give sound change in development of a theory of speech perception. We presume that synchronic variation precedes and is a precursor of full-fledged sound change. For example, English speakers may differ in the extent to which vowels are nasalized before nasal consonants, depending at least in part on such factors as speech rate. Variation of this sort—partially contextual and partly speaker dependent—could easily have been included in the sources of interspeaker variation that, as Rose *et al.* note, can cause difficulties for ASR. For Ohala, listener-based sound changes like the sporadic vowel nasalization in Hindi rely on listeners' misparsing the acoustic signal, in the Hindi case falsely attributing the spectral effects on a vowel of an adjacent voiceless fricative to an absent nasal; in his view, this misattribution can only occur if listeners are perceiving spoken language in terms of acoustic patterns like that produced by an adjacent nasal or voiceless fricative. Fowler accepts Ohala's specific account of the Hindi nasalization, but rejects his inference based on this account; in her view, the spontaneous nasalization could not have occurred if some listener had not perceived the lowered velum of a nasal instead of the glottal opening and/or increased airflow of a voiceless fricative. Further, she suggests that sound change might not, in fact, be relevant to a theory of speech perception. In her view, constraints on the adequacy of the acoustic signal as information for the vocal tract gestures that shaped it need not be built into the linguistic system that the acoustic speech signal carries; in fact, to do so would be to confuse the signal with that which it signals.

While the papers that directly address the human perception problem only touch on ASR, and the Rose *et al.* paper does not address human perception, there are further bases for comparison. It is reasonable to assume that including more knowledge about the source of speech can aid ASR in terms of accuracy; whether or not this improved accuracy is worth the cost is another question. As Rose *et al.* point out, using simultaneous vocal tract images and speech waveforms can improve speech recognition results. Fowler says that perception of gestures is not logically necessary; rather, it is biologically necessary. Fowler is claiming that speech perception must be of gestures, because "...perceptual systems were shaped by natural selection to serve the function of acquainting perceivers with components of their niches... ." Thus, while the authors concerned with ASR invoke engineering pragmatics, Fowler invokes natural selection. Engineering design does not necessarily completely mimic what has evolved in nature; as both Ohala and Lindblom point out, airplanes do not flap their wings. However, what evolves in nature and what is designed by engineers depend on what each domain starts with. (Nature did not have internal combustion engines for propulsion during flight when some animals began to fly.) Rose *et al.* are grappling with the issue of engineering feasibility: Have our systems and understanding reached the point where we can include articulatory knowledge into speech recognition algorithms at a reasonable cost? It may be that nature and engineering will turn out to converge on this issue, but this is only because they have similar tools.

Adaptation on evolutionary time scales certainly plays a role for Fowler, because direct perception is biologically necessary in her view. For Lindblom, adaptation is a very important aspect of speech production. He points to instances where talkers adapt so that listeners understand in different circumstances. In Lindblom's view, the talker must insure that the auditory form of an utterance is sufficiently distinct for listeners to understand it. He relates this adaptive behavior to sound change in language. As previously noted, Ohala argues that listeners base their own articulations on at least some aspects of speakers' acoustic productions. This adaptive behavior can cause sound change, because different articulators can be used to produce essentially the same sound. In contrast, there does not appear to be adaptive use of articulatory knowledge in ASR as proposed by Rose, Schroeter, and Sondhi. Rather than adapting the articulatory knowledge brought to bear on the ASR problem to particular talkers, they propose to smooth the representation of the spectral envelope of sampled speech so as to do away with speaker-dependent characteristics. Further, there is no proposal that the articulatory knowledge be adapted to handle intraspeaker variability, environmental noise, or channel variability.

The authors of the critiques provide excellent and valuable commentary on the longer papers. While the authors of the papers and critiques in this section may not agree on all fundamental issues, it becomes clear that there are important parallels between ASR and speech production research. In particular, the difficulty of incorporating knowledge of articulation into ASR parallels the difficulty of determining the importance of articulation in human perception. Dialog of the sort represented in this section should help refine our approaches to the phenomena that compel our interest.

ACKNOWLEDGMENT

This work was supported by NIH Grants No. HD-01994 and No. DC-01247 to Haskins Laboratories.

Faber, A. (1992). "Articulatory variability, categorical perception, and the inevitability of sound change," *Explanation in Historical Linguistics*, edited by G. W. Davis and G. K. Iverson (Benjamins, Amsterdam), pp. 59–75.

Hock, H. H. (1986). *Principles of Historical Linguistics* (de Gruyter, Berlin).

Kiparsky, P. (1988). "Phonological change. Linguistics: The Cambridge Survey," *Linguistic Theory: Foundations*, edited by F. J. Newmeyer (Cambridge U.P., Cambridge), Vol. I, pp. 363–415.

McMahon, A. M. S. (1994). *Understanding Language Change* (Cambridge U.P., Cambridge).

Nelson, M., and Bourlard, H. (1995). "Continuous speech recognition," IEEE Signal Process. Mag., pp. 25–42 (May 1995).