114/2.

SPEECH RESEARCH AT HASKINS LABORATORIES

Carol Fowler and Katherine S. Harris
Haskins Laboratories

INTRODUCTION

In this short history of Haskins Laboratories we have emphasized the abiding themes of the Laboratories' work, and the lines of research for which they are best known, although this tends to shift the emphasis towards older studies and is done at the expense of particular investigations. This focus leaves us without a specific spot to emphasize Haskins' most significant contribution: It has been for about fifty years a place where a substantial group of investigators with expertise in different aspects of speech and language production and perception could work together, in an environment with substantial and often innovative technical resources.

BEGINNING

Phonetics research began at Haskins Laboratories (founded in 1935 in New York City, moved to New Haven, CT, in 1970) late in the Second World War as part of a program headed by FRANKLIN COOPER and ALVIN LIBERMAN, to develop aids for the recently war-blinded, in particular, a reading machine. The machine was built to produce distinct, discrete, and invariant non speech sounds each corresponding to a textual symbol. That is, an acoustic alphabet replaced the written one. When sounds were sequenced at rates fast enough to be within normal reading rates, listeners were unable to determine the order in which the sounds occurred; the rates apparently exceeded those at which the auditory system can resolve the temporal order of discrete sounds. This finding led them to ask why speech perception can occur at rates so much faster than the rates at which users of the reading machine failed. Using the sound spectrograph, then recently developed at Bell Laboratories and its complement, the Pattern Playback, (invented and built at Haskins by Franklin Cooper), Liberman, Cooper and PIERRE DELATTRE began to study the acoustic speech signal, to determine how it represents the consonants and vowels of spoken words, and to discover the acoustic structure (the "cues") essential for their identification by listeners. Spectrographic displays provided a visible indication of the structure in the acoustic speech signal; the Playback transformed either a photographic negative of a spectrogram, or, more usefully, a hand-drawn, often highly schematic, spectrogram into sound. By selectively including and eliminating elements of acoustic structure, Liberman and his colleagues could determine what bits of structure provided information for the different phonetic properties of spoken words.

They made a number of important discoveries (e.g., DELATTRE, LIBERMAN, COOPER and Gerstman, 1952; Liberman, Delattre and Cooper, 1952). Discovery of the acoustic cues for consonants and vowels, among other consequences, enabled development of rules for speech synthesis, first by Frances Ingemann (Ingemann and MERMELSTEIN, 1975) and subsequently, implemented as a computer program, by IGNATIUS MATTINGLY (1968). This advance, among its many other valuable uses, was a crucial step making possible the development of the modern reading machine. A prototype machine was built at Haskins in a project worked on by Cooper, Patrick Nye, Frances Ingemann, Jane Gaitenby and George Scholes, but loss of funding prevented its full implementation.

MOTOR THEORY

The group realized early that speech is not represented acoustically by an alphabet; rather, owing to coarticulation in speech production, information for consonants and vowels in a word is interleaved and is highly context-sensitive. Two early findings were especially important. LIBERMAN et al. (1952) and Carol Schatz (1954) found that the same piece of acoustic structure (a burst centered at 1440 Hz) was heard as /p/ in the context of some following vowels but as /k/ in the context of others. This finding held whether natural subphonemic bits of speech were reassembled from taped natural speech, or whether highly schematized patterns were synthesized on the Pattern Playback. Complementary to this finding, two very different transitions of the second formant (a high rising transition in the context of /i/ and a low falling transition in the context of /u/) specified invariant sounding /d/s to listeners (Liberman, DELATTRE, COOPER and Gerstman, 1954).

These two findings were central to the development of a novel theory of speech perception. Both findings appear to show that the listener's percept bears a closer correspondence to the gestures that produce consonants than to the acoustic signal. Due to coarticulation, the signal provides different "cues" to the same phonetic properties in different phonetic contexts. Due to coarticulation, the only way to get a burst centered at 1440 Hz before the vowel /i/ is to close the lips; the only way to get the same burst before /a/ is to make a velar constriction. Due to coarticulation, the same alveolar constriction produces a high rising second formant transition in the context of coarticulated /i/ and a low falling one in the context of coarticulated /u/. Despite coarticulation, talkers articulate a given consonant in roughly the same way in all phonetic contexts.

The early version of LIBERMAN's motor theory (Liberman, COOPER, Harris and MACNEILAGE, 1962) proposed that the invariant percepts correspond to invariant articulations but variable acoustic signals develop when listeners learn to associate sensory feedback from their speech musculature to the consequent acoustic signals. The motor theory stimulated considerable phonetics research at Haskins Laboratories and elsewhere.

SPEECH AS A BIOLOGICAL SPECIALIZATION

Although LIBERMAN initially supposed that the motor percept was achieved by a process of associative learning, discussions with scientists who studied the communication systems of other animals helped him to develop a different idea (Liberman, COOPER, Shankweiler and STUDDERT-KENNEDY, 1967). The communication systems of other animals are evolutionary

achievements. Further, some other animals (certain birds, for example) had been found to recruit their neural systems for call production during the act of perception (Nottebohm, 1975), not unlike what Liberman proposed that human speech perceivers do. Liberman and colleagues proposed, now, that phonetic perception is no less a biological specialization of humans than is the ability to produce and understand syntactically structured sentences. One consequence of this idea was research conducted by Donald Shankweiler, Michael Studdert-Kennedy and BRUNO REPP (Studdert-Kennedy and Shankweiler, 1970; Repp, 1975) using dichotic listening to test for a left-hemisphere specialization for speech perception. Highly reliable findings were that stop consonants especially led to a strong right ear advantage in dichotic listening. (Due to the right ear's privileged access to the left hemisphere, these findings have been interpreted as evidence for left hemisphere superiority in stop perception.)

A second consequence of the idea that speech is a biological specialization of humans was a search for perceptual evidence for a distinct module of the brain (in Fodor's (1983) sense of "module") sub-serving speech perception (and production). The most compelling evidence was the finding of "duplex perception" obtained originally by Timothy Rand (1974) at Haskins but pursued further, among others, by LIBERMAN, Douglas Whalen and Virginia Mann (Mann and Liberman, 1983; Whalen and Liberman, 1987). Duplex perception is observed when all of a consonant-vowel syllable except for its third formant transition is presented to one ear, while the missing transition is presented to the other ear. Remarkably, listeners integrate the acoustic structure across the ears, hearing a syllable the identity of which is determined by the transition, but the speech-likeness of which depends on integration of the transition with the remainder of the syllable. Perception is duplex, because the listener also hears a chirp-like sound in the ear to which the isolated transition was presented; the chirp is the transition heard by itself. The important finding is that listeners hear the transition in two ways simultaneously, as part of a speech syllable and by itself. This suggests that two perceptual systems are being engaged by the stimulation; if so, one may be a special-to-speech system that renders a syllabic percept, and the other is likely to be a general auditory system that renders an apparently literal percept of the frequency glide that constitutes the transition.

ACOUSTIC CUES FOR SPEECH

While these theoretical developments were important in motivating work at the Laboratories, systematic experiments on the nature of the cues for speech perception continued, and suggested ways of tying the cues to speech perception to the details of speech production. Work by ARTHUR ABRAMSON and LEIGH LISKER showed the importance of the acoustic consequences of the timing of the burst of energy signaling stop release and of vocal fold pulsing in signaling the distinction between "voiced" and "voiceless" stop consonants. Interestingly, investigations across languages showed that the timing of these events was language specific — that is, the same interval was judged as e.g., /p/ or /b/, depending on the target language, that is, the category boundary moved around (Lisker and Abramson, 1964). Two exciting lines of work resulted. At a perceptual level, Liberman showed that discrimination of speech sounds is good at phonetic category boundaries, but quite poor within them. In languages with different category boundaries, the discrimination peak moves correspondingly (LIBERMAN, Harris, Hoffman and Griffith, 1957). This relationship between speech perception and auditory psychophysics piqued the interest of experimental psychologists in both the phenomenon itself and the study of speech itself as a worthy subject for examination within the framework of general cognitive psychology.

A second development was the interest generated in the articulatory events underlying the acoustic events. Collaboration with the Institute of Logopedics and Phoniatrics of the University of Tokyo allowed Haskins Laboratories to examine events at the larynx directly with fiberoptic instrumentation, and show that the dynamic differences seen in the acoustic signal mirrored the physiological timing of events at the larynx and upper articulators (Sawashima, ABRAMSON, COOPER and LISKER, 1970). Subsequent research has shown that the study of voice onset time is an important tool in tracing the development of speech from babbling to adulthood, in characterizing the difference between pathological and normal speech, as well as in documenting cross-language differences.

SPEECH PRODUCTION

Katherine Harris and PETER MACNEILAGE began to study speech production, initially to test the motor theory's prediction that EMG signals from muscles used in production of phonetic properties would be found to be less context-sensitive than the corresponding movement signals. The tests failed to support the theory, leading to modifications of it. However, Katherine Harris and, later, her students, Fredericka Bell-Berti, Lawrence Raphael, and Gloria Borden recognized the important puzzles requiring solution for our understanding of speech production, and work in that domain began to develop along lines somewhat independent of research on perception.

The initial interest in the muscles of articulation as a potential source of invariant signals led, eventually, to the development of a novel set of methods for studying them. The researchers from the University of Tokyo brought with them techniques for using fine hooked wires inserted in the articulatory and laryngeal muscles to examine the organization of muscles involved in the articulation of various sounds. Bell-Berti (1976) examined the muscles of velopharyngeal closure. The Tokyo researchers, with Tom Baer, made contributions to laryngeal articulation (Hirose, Lee and Ushijima, 1974; see the review by Harris, 1981). Raphael and Bell-Berti (1975) examined the muscular correlates of the distinction between tense and lax in vowels.

COARTICULATION

One area in which researchers at the Laboratories have made important theoretical and methodological contributions to an understanding of speech production is in the area of coarticulation. Based on their research findings, Bell-Berti and Harris (1979) challenged a popular theory of coarticulation, the look-ahead model, based on feature descriptions of phones. According to that theory, talkers will begin producing a phonetic feature (such as lip rounding or nasalization), as early as they can in a word and therefore during any segments that do not require an opposite-sign feature value (lack of rounding or nasalization). Bell-Berti and Harris found that anticipations are much shorter than previous research had suggested and tend to begin an invariant interval before other phonetic properties of, for

example, a rounded or nasal segment. Demonstrating this involved a methodological improvement over earlier research designs. In the study of rounding, for example, lip movement or lip muscle activity in the phonetic environment of a rounded vowel is evaluated relative to lip movement or lip muscle activity in the same phonetic environment, but of an unrounded vowel. In this way, lip movement that may accompany production of unrounded segments can be factored from coarticulatory lip rounding.

ARTICULATORY PHONOLOGY

Beginning in the late seventies, work on production and perception moved closer together, and related work on a theory of articulatory phonology was initiated. One idea motivating the joint attention to phonology, production, and perception was that independent development of these three domains may have exaggerated apparent incompatibilities among their component units. Phonological theory offered units that could not be implemented nondestructively as vocal tract activity; vocal tract activity, therefore, could not be supposed to cause an acoustic signal that conveyed the phonological units directly; accordingly, perception of speech necessarily was construed as indirect and reconstructive.

One line of theoretical development and experimental work, therefore, was to discover the consequences of assuming that phonological units are, as it were, designed for the vocal tract. Catherine Browman and Louis Goldstein proposed and developed an articulatory phonology (1986) in which primitives of the phonology are articulatory gestures, rather than abstract features or phonemes. The theory has excited considerable discussion and debate in the literature. They have implemented their theory as a Linguistic Gestural Model on the computer that serves as a component of an integrated model of production.

ACTION THEORY AND TASK DYNAMICS

As for production itself, the search began for physical evidence of gestures. Stimulated by theoretical advances in our understanding of intentional action made by Michael Turvey, it was proposed by FOWLER, Rubin, REMEZ and Turvey (1980) proposed that articulatory "units" of action in speech are coordinative structures (that is, synergies). For example, in production of /b/, the two lips and the jaw temporarily form a "special purpose device" the goal of which is to produce a constriction at the lips. The synergies produce the articulatory gestures described by Browman and Goldstein's articulatory phonology. Research by Vincent Gracco (1988) and by Scott Kelso, Betty Tuller and colleagues (Kelso, Saltzman and Tuller, 1986) tested and confirmed proposals that transient coordinations form during speech. Subsequently, these theoretical ideas were made more explicit and were related to dynamical systems theory by Elliot Saltzman (Saltzman and Kelso, 1987), who developed a "task dynamic" model of speech production. Browman and Goldstein's Linguistic Gestural Model provides "gestural scores" for words to Saltzman's task dynamic model. In turn, the task dynamics model controls Haskins' articulatory synthesizer. The synthesizer first developed by PAUL MERMELSTEIN (1973) [at Bell Labs] and further developed by Mermelstein, Pat Nye, Tom Baer and Philip Rubin (Rubin, Baer and Mermelstein, 1981), produces acoustic speech signals from vocal tract shapes and motions. Accordingly, the integrated model produces theory- and model-driven acoustic speech signals.

THE LINK BETWEEN PRODUCTION AND PERCEPTION

An essential characteristic of any communication system is that it must consistently achieve "parity" — here, a relation of equivalence between messages sent and received. A proposal that gestures are units of words as language users know them and are units in actions of the vocal tract producing speech is a proposal that the units of the spoken message are preserved in vocal-tract activity — a parity-fostering characteristic of speech if the proposal is accurate. If listeners perceive gestures, then units of the original message (at the phonological level of description) are preserved throughout a communicative exchange, further fostering achievement of parity. Two theories of perception developed at Haskins Laboratories suggest that listeners perceive gestures, and a recognition model has tested the feasibility of gesture extraction in machine recognition of speech.

One theory is the revised motor theory (LIBERMAN and MATTINGLY, 1985), which proposes that gesture extraction is an important function of a specialized phonetic module of the brain. In that theory, gesture extraction is achieved by recruitment of the speech motor system in perception. The other is direct realist theory (FOWLER, 1986), which claims that listeners perceive gestures because they are what is conveyed by the structure in the acoustic speech signal. As for the recognition model, Richard McGowan (1994) has used genetic algorithms to find a best-fitting gestural solution to acoustic input representing VCVs.

Together, the theories of articulatory phonology, of task dynamics in production and of gesture perception constitute the most integrated account of these components of language use in the field of speech to date.

FIRST LANGUAGE DEVELOPMENT

Finally, researchers at Haskins have made significant contributions to the field of phonetics in their work on development of speech production and perception in infants and young children. The work on production carried out by MICHAEL STUDDERT-KENNEDY and his students (Nittrouer, Studdert-Kennedy and McGowan, 1989; Goodell and Studdert-Kennedy, 1993) has shown that phonological (and therefore gestural) structure in words emerges gradually from larger, syllable like structures. The work on perception carried out by Catherine Best and colleagues (Best, 1994) has led to a theory (the Perceptual Assimilation Model) that explains how the infant's perception of native and nonnative phones undergoes change beginning at about 10 months of age. Nonnative phones that are very similar to native phones assimilate to native categories. Two nonnative phones that assimilate to the same native category become difficult to discriminate by 10 months of age even though they may have been readily discriminated at younger ages. Pairs of nonnative phones that assimilate to different native categories remain discriminable.

CLINICAL RESEARCH

Finally, we should point out that the Laboratories have consistently carried out work on problems of clinical populations in parallel with investigations of normal production and perception. We have performed studies of stuttering (e.g., Story,

Alfonso and Harris, 1996) and the speech of the deaf (e.g., McGarr and Löfqvist, 1982). Most notably, researchers at the Laboratories have developed a theory of phonological awareness as a way of accounting for the difference between good and poor readers (e.g., LIBERMAN and Shankweiler, 1989).

ACKNOWLEDGMENTS

The work of the Laboratories has been supported over many years by the National Institute of Child Health and Human Development and the National Institute of Deafness and Other Communication Disorders of the National Institutes of Health, and the National Science Foundation, among other institutions. Lastly we would like to acknowledge the help, both tangible and intellectual, of Caryl and Edna Haskins.

REFERENCES

BELL-BERTI, F. 1976. An electroencephalographic study of velopharyngeal function in speech. *Journal of Speech and Hearing Research*, 19, 225-240.

BELL-BERTI, F. and K. S. HARRIS. 1979. Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America*, 65, 1268-1270.

BEST, C. T. 1994. Learning to perceive the sound pattern of English. In C. Rovee-Collier and L.P. Lipsitt (Eds.), *Advances in infancy research*. Norwood, NJ: Ablex. 217-304.

BROWMAN, C. P. and L. GOLDSTEIN. 1986. Towards an articulatory phonology. *Phonology Yearbook*, 3, 219-252.

DELATTRE, P. C., A. M. LIBERMAN, F. S. COOPER, and L. J. GERSTMAN. 1952. An experimental study of the acoustic determinants of vowel colour; Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, 8, 195-210.

FODOR, J. A. 1983. *The modularity of mind*. Cambridge, MA: MIT Press.

FOWLER, C. A. 1984. Realism and unrealism: A reply. *Journal of Phonetics*, 11, 303-322.

FOWLER, C. A. 1986. An event approach to a theory of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3-28.

FOWLER, C. A., P. E. RUBIN, R. REMEZ, and M. TURVEY. 1980. Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production*. New York: Academic Press. 373-420.

GOODELL, E. W. and M. STUDDERT-KENNEDY. 1993. Acoustic evidence for the development of gestural coordination in the speech of 2 year olds: A longitudinal study. *Journal of Speech and Hearing Research*, 36, 707-727.

GRACCO, V. L. 1988. Timing factors in the coordination of speech movements. *Journal of Neuroscience*, 8, 4628-4634.

HARRIS, K. S. 1981. Electromyography as a technique for laryngeal investigation. In C.L. Ludlow and M.O.C. Hart (Eds.), *Proceedings of the conference on the Assessment of Vocal Pathology. ASHA Reports 11*. Rockville, MD: American Speech-Language-Hearing Association. 70-87.

HIROSE, H., C. Y. LEE, and T. USHIJIMA. 1974. Laryngeal control in Korean stop production. *Journal of Phonetics*, 2, 145-152.

INGEMANN, F. and P. MERMELSTEIN. 1975. Speech recognition through spectrogram matching. *Journal of the Acoustical Society of America*, 57, 253-255.

KELSO, J. A. S., E. L. SALTZMAN, and B. TULLER. 1986. The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, 14, 29-59.

LIBERMAN, A. M., F. S. COOPER, K. S. HARRIS, and P. MACNEILAGE. 1962. A motor theory of speech perception. In *Proceedings of the Speech Communication Seminar*. Stockholm: Speech Transmission Laboratory, Royal Institute of Technology.

LIBERMAN, A. M., F. S. COOPER, D. P. SHANKWEILER, and M. STUDDERT-KENNEDY. 1967. Perception of the speech code. *Psychological Review*, 74, 431-461.

LIBERMAN, A.M., P. C. DELATTRE, and F. S. COOPER. 1952. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 65, 497-516.

LIBERMAN, A. M., P. C. DELATTRE, F. S. COOPER, and L. J. GERSTMAN. 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 68, 8, 1-13.

LIBERMAN, A. M., K. S. HARRIS, H. S. HOFFMAN, and B. C. GRIFFITH. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-268.

LIBERMAN, A. M. and I. G. MATTINGLY. 1985. The motor theory of speech perception revised. *Cognition*, 21, 1-36.

LIBERMAN, I. Y. and D. P. SHANKWEILER. 1989. *Phonology and reading disability: Solving the reading puzzle*. Ann Arbor, MI: University of Michigan Press.

LISKER, L. and A. S. ABRAMSON. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384-422.

MANN, V. A. and A. M. LIBERMAN. 1983. Some differences between phonetic and auditory modes of perception. *Cognition*, 14, 211-235.

MATTINGLY, I. G. 1968. Experimental methods for speech synthesis by rule. *IEEE Transactions on Audio and Electroacoustics*, AU-16, 198-202.

MCGARR, N. S. and A. LÖFQVIST. 1982. Obstruent production in hearing-impaired speakers: Interarticulator timing and acoustics. *Journal of the Acoustical Society of America*, 72, 34-42.

MCGOWAN, R. S. 1994. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 14, 19-48.

MERMELSTEIN, P. 1973. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53, 1070-1082.

NITTROUER, S., M. STUDDERT-KENNEDY, and R. S. MCGOWAN. 1989. The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, 32, 120-132.

NOTTEBOHM, F. 1975. A zoologist's view of some language phenomena with particular emphasis on vocal learning. In E.H. Lenneberg and E. Lenneberg (Eds.), *Foundations of language development*. New York: Academic Press. Vol. 1., 61-103.

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

RAPHAEL, L. J. and F. BELL-BERTI. 1975. Tongue musculature and the feature of tension in English vowels. *Phonetica*, 32, 61-73.

REPP, B. H. 1975. Distinctive features, dichotic competition, and the encoding of stop consonants. *Perception and Psychophysics*, 17, 231-240.

RUBIN, P. E., T. BAER, and P. MERMELSTEIN. 1981. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70, 321-328.

SALTZMAN, E. L. and J. A. S. KELSO. 1987. Skilled actions: A task dynamic approach. *Psychological Review*, 94, 84-106.

SAWASHIMA, M., A. ABRAMSON, F. S. COOPER, and L. LISKER. 1970. Observing laryngeal adjustments during running speech by use of a fiberoptics system. *Phonetica*, 22, 193-201.

SCHATZ, C. 1954. The role of context in the perception of stops. *Language*, 30, 47-56.

STORY, R. S., P. F. ALFONSO, and K. S. HARRIS. 1995. A pre- and post-therapy comparison of the kinematics of fluent speech of persons who stutter. *Journal of Speech and Hearing Research*, 39, 991-1005.

STUDDERT-KENNEDY, M. and D. P. SHANKWEILER. 1970. Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, 48, 579-594.

WHALEN, D. H. and A. M. LIBERMAN. 1987. Speech perception takes precedence over nonspeech perception. *Science*, 237, 169-171.