XXI. SPEECH COMMUNICATION[*]

Prof. K. N. Stevens	N. Benheim	J. S. Perkell
Prof. M. Halle	T. H. Crystal	Eleanor C. River
Prof. J. M. Heinz	H. J. Hebert	A. W. Slawson
Dr. Paula Menyuk	Jane A. Heinz	R. S. Tomlinson
Dr. S. E. G. Öhman[†]	W. F. Henke	J. T. Williams
	Y. Kato[‡]	

RESEARCH OBJECTIVES

The objectives of our work are to further our understanding of: (a) the process whereby human listeners decode an acoustic speech signal into a sequence of discrete linguistic symbols such as phonemes, and (b) the process whereby human talkers encode a sequence of discrete linguistic symbols into an acoustic signal.

During the past year we have continued studies of the acoustic and articulatory processes involved in speech production in an attempt to gain further understanding of the relations between events at these two levels and to account for the contextual influences that occur when sequences of speech sounds are generated. These studies have included analyses of cineradiographic films and measurements of poles and zeros associated with the spectra of vowels and some consonants in various contexts. We have also been working on the development of a new terminal-analog speech synthesizer that will be controlled from a digital computer, and a real-time spectral analysis system that will provide spectral inputs in quantized form to a digital computer. During the forthcoming year we expect to complete the real-time input system and the terminal-analog synthesizer, and to commence a program of research on speech synthesis and speech perception with the synthesizer. The analysis of cineradiographic films will continue, and we shall explore the possibilities of using a digital computer for the storing, analysis, and display of the data derived from frame-by-frame tracings made from these films. Other projects that are being initiated are the development of a model for laryngeal activity in speech and a model for describing the shapes and motions of the vocal tract during speech production. During the past months we have begun studies of speech development in children, and the first projects in this area will include studies of speech perception in children and acoustic analysis of the speech of children.

K. N. Stevens, M. Halle, J. M. Heinz

A. STUDIES OF THE DYNAMICS OF SPEECH PRODUCTION

During the past months an expanded detailed analysis of lateral cineradiographs of the vocal tract has been undertaken. A preliminary analysis of this type and a description of the film that has been used have been presented previously.[1,2] The objective of the expanded work is to derive a body of information about the dynamic behavior of the vocal tract in order to better understand its physiology and to aid in the formulation of a

[*] This work is supported in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, U.S. Air Force, under Contract AF19(628)-3325; and in part by the National Science Foundation (Grant GP-2495), the National Institutes of Health (Grants NB-04332-02 and MH-04737-04), and the National Aeronautics and Space Administration (Grant NsG-496).

[†] On leave from the Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden.

[‡] On leave from the Nippon Electric Company, Limited, Tokyo, Japan.

general set of rules governing the encoding of linguistic signals into acoustic signals.

The film that was originally described has been improved in quality with the use of the logEtronic automatic dodging process. This process involves the use of a logEtronic printer which yields a copy of the original film with increased contrast and definition. The Satellite Tracking Service of the U. S. Weather Bureau has been helpful in processing the film on their printer, and their cooperation is gratefully acknowledged.

From the improved film, frame-by-frame tracings of a number of utterances have been made. The tracing process has been improved and speeded up with the use of overlays corresponding to the structures that change very little in shape during speech production. These overlays, representing the vertebrae, upper jaw, and lower jaw are the bases for coordinate systems used in describing the positions and shapes of movable vocal-tract structures. The coordinate systems have been devised so that the dynamic information will be as meaningful and consistent as possible with known anatomical and physiological constraints. Thus over the 30-odd frames covering a typical utterance the movements of the vertebrae, maxilla, mandible, tongue, soft palate, hyoid bone, laryngeal structures, and anterior pharyngeal wall (dorsum of the tongue) can be described with varying degrees of accuracy. Spectrograms made from a simultaneous tape recording of the text spoken for the film allow close correlation of acoustical and articulatory events. From this correlation it is hoped that physiological features may be determined to correspond with phonetic features of speech.

Position versus time plots have been made for seven utterances that have the form /hə'Cɛ/, where /C/ represents the consonants /t, d, s, z, n, k, p/. By superimposing plots of one parameter for the several consonants, features of dynamic behavior can be distinguished and correlated with phonetic features of the phoneme in this particular context.

An example of the type of comparison which can be made is shown in Fig. XXI-1, which is a temporal plot of pharynx width (distance from the vertebrae to the dorsum of the tongue) for /hə'dɛ/ and /hə'tɛ/. The two curves are so aligned that the release of the consonants coincides. The mark on each curve to the left of this time origin indicates the time at which consonant closure occurs, and the mark on the /hə'tɛ/ plot at ~50 msec represents the time when voicing begins. It can be seen that during the closure the pharynx width for the voiced stop increases appreciably relative to that for the voiceless stop. The increase in volume of the pharynx during the voiced stop seems to be approximately equal to the volume increase that would be necessary to allow air to be expelled into the vocal tract to permit vocal-cord vibration. The spectrogram for /d/ shows voicing that dies off toward the end of the stop gap, as the expansion of the vocal tract possibly reaches a limit for this particular configuration.

This type of explanation suggests a physiological basis for the term "lax" (versus) "tense") as applied to /d/ and /t/. The "tense" vocal-tract configuration for /t/ would

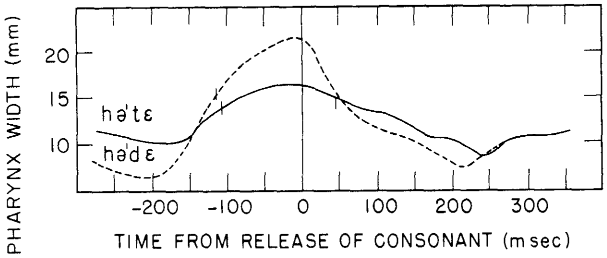

Fig. XXI-1. Pharynx width vs time for /hə'tɛ/ and /hə'dɛ/. Pharynx width is measured as the horizontal distance from the anterior surface of the bodies of the vertebrae forward to the dorsum of the tongue at the level of the bottom of the body of the third cervical vertebra. All curves are aligned temporally with respect to release of the consonant as determined from the spectrogram.

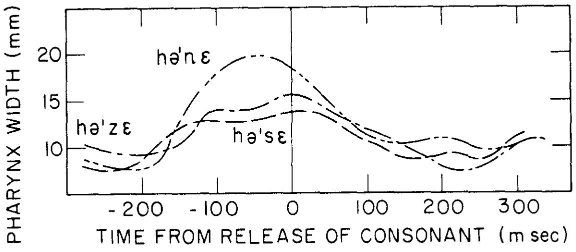

Fig. XXI-2.

Same measurement as in Fig. XXI-1 for /hə'sɛ/, /hə'zɛ/, and /hə'nɛ/.

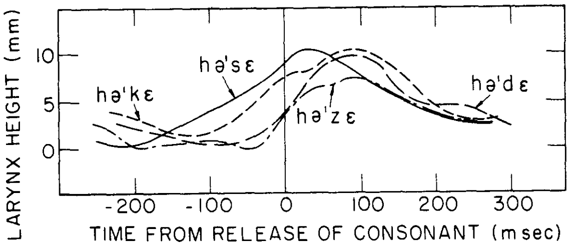

Fig. XXI-3.

Larynx height vs time for /hə'kɛ/, /hə'sɛ/, /hə'dɛ/, and /hə'zɛ/ measured as the height of the posterior end of the vestibule from an arbitrary plane at the level of the seventh cervical vertebra.

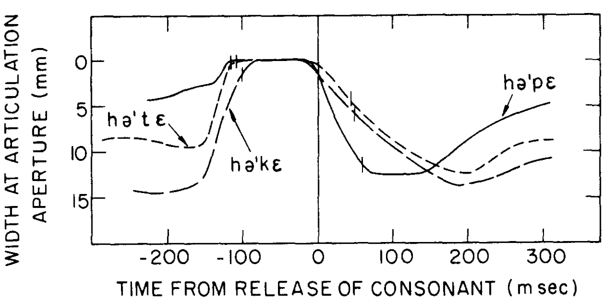

Fig. XXI-4. Width at aperture of articulation vs time for /hə'tɛ/, /hə'kɛ/, and /hə'pɛ/ measured along a line approximately perpendicular to the articulatory movement for the tongue tip in /t/, the body of the tongue in /k/, and the lips in /p/.

imply a rigid vocal-tract wall, which would not expand to permit the increase in volume needed for a voiced stop. Presumably, a similar "tense" configuration exists for the voiceless unaspirated stop consonants occurring in certain languages (and in English after /s/). For such stop consonants an instruction to the larynx musculature to assume a configuration appropriate for voicing would not result in vocal-cord vibration until release of the stop, whereas a "lax" vocal-tract configuration would permit a limited amount of air to pass through the glottis, with consequent glottal vibration.

Figure XXI-2 shows a similar, but less pronounced difference in pharynx width behavior for /z/ and /s/. In the case of the voiced nasal continuant /n/, there is also a widening of the pharynx, but presumably for different reasons. The articulation of /n/ is different from that of /d/ and /z/ in two respects: First, there is no pressure build-up which must be counteracted by tongue or pharynx musculature during /n/; second, the palatopharyngeus muscle (which helps stabilize the position of the soft palate in its nasalized consonant position) would influence the shape and rigidity of the pharynx wall.

The necessity for larger supraglottal volume may partly explain the tendency toward lower larynx position for the voiced consonants during the consonant only, as shown in Fig XXI-3, which is a plot of larynx height against time for several consonants.

As a second example, in Fig. XXI-4 there is a comparison of the temporal change in midsaggital vocal-tract width at the place of articulation in /hə'pɛ/, /hə'tɛ/, and /hə'kɛ/. The three curves are aligned with respect to point of consonant release (as determined from the spectrogram), and are plots of distance from tongue tip to palate in /t/, distance between the lips in /p/, and distance from the top of the tongue to the junction of hard and soft palates in /k/. The small vertical line previous to release indicates onset of closure, and the vertical line following release indicates the onset of voicing. This similar dynamic behavior for three very different sets of articulatory muscles suggests the concept of a common mechanism of control. It is apparent that a number of complex factors contribute to this grossly similar behavior. There must be interrelationships between the observed features of the dynamic behavior (e.g., the speed of articulation), and obvious differences in anatomy, physical and physiological properties of the structures involved, and the methods of measurement.

Figure XXI-3 (larynx height) demonstrates a gross behavior which upon reinspection of the film proved to be primarily due to vocalization of the entire word rather than to a particular word segment. During two sentences that form part of the film and that have been examined informally, the larynx seems to rise to a vocalizing position and remain there until end of a phrase. Differences in detailed behavior of the curves can, however, be interpreted in terms of features present in the utterance, as noted above.

At present, the large amount of data obtained from these first seven utterances is still under analysis. Also the same type of data is being taken for several vowels in the same consonantal environment. Correlation of these two groups of information should help in distinguishing certain general physiological concepts that will be useful in establishing guidelines for further uses of these techniques.

J. S. Perkell

References

1. K. N. Stevens, Studies of the dynamics of speech production, Quarterly Progress Report No. 71, Research Laboratory of Electronics, M.I.T., October 15, 1963, pp. 203-205.

2. K. N. Stevens and S. Öhman, Cineradiographic Studies of Speech, Speech Transmission Laboratory, Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, Sweden, July 15, 1963, pp. 9-11.