XVI. SPEECH COMMUNICATION

Academic and Research Staff

Prof. K. N. Stevens	Dr. W. L. Henke	Dr. J. S. Perkell
Prof. M. Halle	Dr. A. W. F. Huggins	Dr. R. A. Stefanski
Dr. Margaret Bullowa	Dr. A. R. Kessler	Dr. Jacqueline Vaissière**
Dr. R. Carré*	Dr. D. H. Klatt	Mary M. Klatt
Dr. K. Hashimoto†	Dr. Paula Menyuk‡	Estill Putney

Graduate Students

T. Baer	Ursula Goldstein	S. Maeda
J. C. Bernstein††		B. Mezrich

RESEARCH OBJECTIVES AND SUMMARY OF RESEARCH

1. Timing Studies

National Institutes of Health (Grant 2 RO1 NS04332-11)

A. W. F. Huggins, D. H. Klatt

A major weakness in our present understanding of speech production is the lack of a theory, or even a description, of how speech timing is organized. A complete model of the speech-production process should include rules to specify the durations of phonetic segments, syllables, words, stress groups, and phrases. The discovery of appropriate rules for English and for other languages has proved to be a difficult undertaking because of the number and variety of low-level adjustment rules that express universal or language-specific durational differences between segments, and durational interactions between adjacent segments.

Major gaps still remain in our ability to describe timing rules for the production of an utterance, and how the rules are organized. For example, it is still not known whether speakers of English behave as if a basic temporal framework for an utterance is determined by sentence syntactic structure and stress pattern, and individual words are then lengthened or shortened in order to fit into the specified timing framework. On the basis of perceptual impressions, English has often been called stress-timed, but this may be due to other factors such as phonological structure, and there may not be a tendency toward local adjustment of segmental durations in order to satisfy constraints such as equal stress timing. Another unresolved issue concerns the influence of semantic information on the duration of a word in a sentence. As a final example, there are now conflicting data on the durational effect of adding syllables to a root or base form. Lindblom,[1] Barnwell,[2] and others have observed that syllables in the base become shorter as more syllables are added to the word. Klatt[3] observed shortening only for syllables added to the end of a word, and Harris and Umeda[4] found no shortening on

---

*On leave from École Nationale Supérieure d'Électronique et de Radioélectricité, Grenoble, France.

†On leave from University of Electrocommunications, Tokyo, Japan.

‡Also Professor of Special Education, Boston University.

**Also Instructor of Phonetics, Department of French, Wellesley College.

††Rackham Prize Fellow from University of Michigan, 1973-74.

the average in spoken sentences. Huggins[5] has shown, however, that under some circumstances (perhaps syntactically determined) shortening of a stressed base syllable occurs even when a following added unstressed syllable does not fall in the same word, thereby implicating levels of organization higher than the word. More experiments are now being performed to explore this finding in more detail.

Other current studies are directed at the measurement of changes to the duration of certain phonetic segments as a function of speaking rate, and the measurement of phrase, word, and segment durations in noun phrases composed of monosyllabic words, where these noun phrases are in various sentence contexts. Fundamental frequency (F_o) contours within these utterances will also be studied. An attempt will be made to organize the data in terms of simple rules that specify the conditions under which a rising, a falling, or a steady F_o contour occurs on a syllable. Similar work for French sentence material is also in progress.

Our primary goal is to develop a descriptive theory of segmental duration in English. The ultimate objective will be to understand how the productive, perceptive, and memory capabilities of a human speaker (or of a child who is faced with the problem of making sense out of the acoustic patterns of speech) impose constraints on the allowed temporal attributes of speech. Preliminary attempts to specify one form for such a theory[6] have led to a better understanding of the issues involved and the data needed. Now that several theories have been proposed for other languages, it would seem to be the time to examine these theories closely and attempt to measure their descriptive adequacy for English.

References

1. B. E. F. Lindblom, "Temporal Organization of Syllable Production," Speech Transmission Laboratory QPSR No. 2-3, Stockholm, Sweden, 1-5 (1968).

2. T. P. Barnwell III, "An Algorithm for Segment Duration in a Reading Machine Context," Technical Report 479, Research Laboratory of Electronics, M. I. T., January 15, 1971.

3. D. H. Klatt, "Vowel Duration as a Function of the Syllabic Structure of a Word," J. Acoust. Soc. Am. 54, 312-313 (A) (1973).

4. M. S. Harris and N. Umeda, "Parametric Analysis of Vowel Duration in Single and Multi-syllable Words," a paper presented at the 86th Meeting of the Acoustical Society of America, Los Angeles, California, November 1, 1973.

5. A. W. F. Huggins, "On Isochrony and Syntax," in G. Fant and L. Chistovich (Eds.), Proc. Symposium on Auditory Analysis and Perception of Speech, Leningrad, U. S. S. R., August 1973 (to be published in Supplement to the Journal of Phonetics).

6. D. H. Klatt, "Generative Theory of Segmental Duration in English," J. Acoust. Soc. Am. 51, 101 (A) (1972).

7. B. E. F. Lindblom, "Some Temporal Regularities of Spoken Swedish," in G. Fant and L. Chistovich (Eds.), Proc. Symposium on Auditory Analysis and Perception of Speech, Leningrad, U. S. S. R., August 1973 (to be published in Supplement to the Journal of Phonetics).

2. Perception of Speech and Speechlike Sounds

National Institutes of Health (Grant 2 RO1 NS04332-11)

A. W. F. Huggins, D. H. Klatt, K. N. Stevens

a. Rapid Spectrum Change

Efforts are under way to determine the role of rapid spectrum change in the perception of speech.[1] One line of research involves the psychophysics of brief synthetic stimuli that resemble the kinds of rapid spectral changes that occur in natural stop-vowel sequences.[2]

For example, it has been determined experimentally that stimuli analogous to the mirror-image formant patterns of a vowel-stop and stop-vowel are not judged to be perceptually similar.[3] One implication of this is that the acquisition of a stop-vowel such as "ba" by a child learning English does not imply that the corresponding (mirror-image) acoustic pattern for the vowel-stop "ab" will automatically be heard as the same stop by the child. Postvocalic stops must be learned as essentially new acoustic patterns, although stop categorization might be aided by association with the motor events that occur during the production of VCV utterances by the child.

Similarity judgments among a wide variety of stimuli indicate that the frequency region of the second formant is most sensitive to rapid spectrum change.[4] Experiments are now under way to probe the extent of this effect. In addition, we hope to propose physiological mechanisms that might account for the data, and to estimate the importance of this effect for the perception of speech.

A related project, which has recently been completed,[5] has shown that one cue for voicelessness in an initial stop consonant in English is the absence of a rapid spectrum change at the onset of voicing, following the interval of aspiration.

References

1. K. N. Stevens, "Potential Role of Property Detectors in the Perception of Consonants," Quarterly Progress Report No. 110, Research Laboratory of Electronics, M. I. T., July 15, 1973, pp. 155-168. A revised version will appear in G. Fant and L. Chistovich (Eds.), Proc. Symposium on Auditory Analysis and Perception of Speech, Leningrad, U. S. S. R., August 1973 (to be published in Supplement to the Journal of Phonetics).
2. D. H. Klatt and S. R. Shattuck, "Perception of Chirps that Resemble Rapid Formant Transitions," Proc. Symp. on Auditory Analysis and Perception of Speech (op. cit.).
3. S. R. Shattuck and D. H. Klatt, "Perception of Mirror-Image Acoustic Patterns" (in preparation for publication).
4. D. H. Klatt and S. R. Shattuck, "Perception of Rapid Spectrum Changes in the Frequency Range of the Second Formant" (in preparation for publication).
5. K. N. Stevens and D. H. Klatt, "The Role of Formant Transitions in the Voiced-Voiceless Distinction for Stops" (to appear in J. Acoust. Soc. Am.).

b. Feature Detectors: Evidence from Adaptation Studies

Evidence from a variety of experimental sources has suggested that human listeners analyze speech sounds into sets of quasi-independent distinctive features at

some stage of the listening process. The existence of feature-extracting mechanisms can be inferred from analyses of perceptual confusions,[1,2] errors in short-term recall,[3,4] response patterns of infants,[5] and other sources.

More recently, a research technique known as selective adaptation has been used to demonstrate the existence of auditory feature detectors in the brain.[6] The method that is employed uses an adapting stimulus to fatigue a particular feature detector and thus shift the phoneme boundary in an identification test involving a set of stimuli lying along an acoustic continuum, e.g., a voice-onset time continuum. Current research (carried out by graduate students in the Department of Psychology, M.I.T., utilizing the facilities of our group) is directed at obtaining a better understanding of the adaptation phenomenon. Nonspeech control stimuli have been used to determine whether adaptation is a peripheral or a central process.[7] Other experiments in progress are examinations of the generality of the adaptation effect in various types of counterbalanced experimental designs and the influence of adaptation on speech production.[8]

References

1. G. A. Miller and P. E. Nicely, "Analysis of Perceptual Confusions among Some English Consonants," J. Acoust. Soc. Am. 27, 338-353 (1955).
2. D. Shankweiler and M. Studdert-Kennedy, "Identification of Consonants and Vowels Presented to Left and Right Ears," Quart. J. Exptl. Psychol. 19, 59-63 (1967).
3. W. A. Wickelgren, "Distinctive Features and Errors in Short-Term Memory for English Consonants," J. Acoust. Soc. Am. 39, 388-398 (1966).
4. D. H. Klatt, "Structure of Confusions in Short-Term Memory between English Consonants," J. Acoust. Soc. Am. 44, 401-407 (1968).
5. P. D. Eimas, E. Siqueland, P. Jusczyk, and J. Vigorito, "Speech Perception in Infants," Science 171, 303-304 (1971).
6. P. D. Eimas and J. D. Corbit, "Selective Adaptation of Linguistic Feature Detectors," Cognitive Psychol. 4, 99-109 (1973).
7. A. E. Ades, "Some Effects of Adaptation on Speech Perception," Quarterly Progress Report No. 111, Research Laboratory of Electronics, M.I.T., October 15, 1973, pp. 121-129.
8. W. E. Cooper, "Perceptuo-Motor Adaptation to a Speech Feature" (in preparation for publication).

c. Auditory Processing of Speech

The intelligibility of temporarily segmented speech (continuous speech broken up into chunks of arbitrary duration by the insertion of silence intervals of arbitrary duration) is related to the durations of the speech and the silence intervals in a way that has interesting implications.[1] Experiments are in progress to explore these earlier results in more detail, and to relate them to psychophysical data, for example, on temporal integration.

References

1. A. W. F. Huggins, "Second Experiment on Temporally Segmented Speech," Quarterly Progress Report No. 106, Research Laboratory of Electronics, M.I.T., July 15, 1972, pp. 137-141.

3. Speech Production

National Institutes of Health (Grant 5 RO1 NS04332-11)

M. Halle, W. L. Henke, D. H. Klatt, K. N. Stevens

Our research on speech production and speech acoustics has two broad objectives. First, we would like to determine relations between positions or movements of the speech-production mechanism and the properties of the acoustic output, and thereby to establish the articulatory positions and timing patterns that give rise to acoustic outputs with distinctive attributes. Such studies could lead to a systematic procedure for predicting the inventory of phonetic categories available for use in language. A second aim is to contribute to an understanding of how articulatory movements are controlled in running speech.

a. Larynx Mechanisms

We are studying laryngeal mechanisms from several points of view. In one project, we are examining in some detail the vibration patterns of excised dog larynges,[1] and the measurements are providing quantitative data on the trajectories of points on the vocal cords during a vibratory cycle; the phase differences of the movements of different points along the medial and superior surfaces of the vocal cords; the conditions under which vibration occurs; and certain physical properties of vocal-cord tissue. A different aspect of our laryngeal studies is an attempt to develop theoretical procedures, which are consistent with experimental observations, for predicting certain modes of laryngeal operation during vowel and consonant production, and hence to establish a basis on which to specify phonetic categories associated with different laryngeal gestures. The theory of larynx behavior is based on the work of other investigators, and is influenced by the experimental findings noted above. Our own experimental studies of human larynx behavior are based, at present, on observations of the waveform and spectrum of the speech signal for speech sounds in English and in other languages in which various modes of laryngeal behavior occur, and on observation of the signal from a small accelerometer attached to the outer surface of the throat. For example, we are finding distinctive types of onset characteristics as the vocal cords begin to vibrate after a voiceless consonant.

References

1. T. Baer, "Measurement of Vibration Patterns of Excised Larynxes," Quarterly Progress Report No. 110, Research Laboratory of Electronics, M. I. T., July 15, 1973, pp. 169-176.

b. Speech Imitation by a Mynah Bird

Birds employ a specialized organ, the syrinx, to produce most of their vocalization. Spectrographic comparisons of the speech imitations of a mynah bird and the speech of its trainer indicate that the bird can produce excellent approximations of most of the acoustic features of human speech. The detailed anatomy of the mynah syrinx has been examined in an attempt to determine the mechanisms of speech mimicry. Similiarities have been noted between the external labia in the syrinx and the vocal cords of the human larynx. This work has led to a general theory of sound generation in the mynah and in other songbirds. Aspects of this theory will be tested in several future physiological studies of bird preparations, and wherever possible the results will be related to the physiology of speech production in humans.

c. Model of the Tongue

Research aimed at understanding the control of speech production is centered, at present, on a project in which a dynamic model of the tongue is being simulated. This model represents the tongue volume by a series of points located within the tongue body, as well as on its surface. These points are held together and activated by elements that simulate connective tissue and muscle. The behavior of this model will be examined and compared with tongue movements observed from cineradiographic studies of speech, and with electromyographic data reported by others. One specific aim of this project is to gain some insight into the temporal pattern of control signals that are required to cause appropriate tongue movements during the production of certain simple speech-sound sequences. In addition, the behavior of the model should be useful in exploring some of the physical properties of the production mechanism as limiting factors in the sequencing of articulatory gestures.

4. Speech Synthesis

National Institutes of Health (Grant 2 RO1 NS04332-11)

D. H. Klatt

a. Synthesis by Rule of an Idiolect

The research area of speech synthesis by rule is at a stage where fairly intelligible speech can be produced from a phonetic representation, or even from English text, but the sound quality of the synthesis is unnatural and machinelike. In an effort to improve the rules for formant specification in a terminal analog speech synthesizer, research has begun on the detailed comparison of formant tracks generated by a set of rules and formant tracks produced by linear prediction analysis of natural sentences read by a single speaker. With this analysis-by-synthesis paradigm it should be possible to optimize tables of formant-frequency targets, improve segmental coarticulation rules, quantify the effects of the sentence stress pattern on formant undershoot, and determine the phonetic representation of sentences produced by the speaker. It will be of particular interest to see if the underlying phonology of a single idiolect can be derived by using this technique. Other issues to be treated include quantification of the limitations of a rule system that manipulates formant parameters directly instead of controlling an articulatory model of the speech-production apparatus.

5. Speech Development and Pathologies

National Institutes of Health (Grant 2 RO1 NS04332-11)

Margaret Bullowa, Paula Menyuk, K. N. Stevens

Ongoing activities related to language development include studies of the organization of body and speech rhythms during communication between infants and adults, based on methods used in ethological field studies,[1] and analysis of the vocalizations of children of preschool and early school ages, particularly the errors made in the production of speech sounds in one-word utterances. (The latter work is being carried out by Professor Menyuk at Boston University, with some collaboration with our group.) Some preliminary experiments on the effectiveness of certain visual displays as speech-reading aids are being planned and, as a first step in this endeavor, displays are being developed of voicing in speech with memory up to a few hundreds of milliseconds (i. e., with a continuously updated display of presence or absence of voicing over this interval).

References

1. Margaret Bullowa, "When Infant and Adult Communicate, How Do They Synchronize Their Behaviors?," prepared for Conference on Face-to-Face Interaction, Chicago, Illinois, August 27-30, 1973.

6. Studies of Interspeaker and Intraspeaker Variability

U. S. Navy Office of Naval Research (Contract ONR N00014-67-A-0204-0069)

W. L. Henke, K. N. Stevens

In seeking universal truths about speech production, one is always faced with the fact that different speakers do not produce the same utterance in an identical fashion and that the same speaker will never produce an utterance in the same way on different occasions. The study of interspeaker and intraspeaker variability can help to provide insight into which attributes of a speech event are invariant across speakers, and how speaker differences can be taken into account in finding these invariant attributes. There is also practical interest in the problem of identifying a speaker or even determining the physiological state of a speaker from measurements on the speech that he produces.

One project in this area includes a study of the formant trajectories of diphthongs and r-colored sounds in English for a number of speakers. These speech events are known to be influenced by dialect differences, and substantial interspeaker differences might also be expected within a given dialect. Preliminary data on several diphthongs support this prediction: There are appreciable differences among speakers in the beginning and end points of the trajectories of the first two formants.

We are also examining individual speaker characteristics by measuring long-term average spectra and distributions of fundamental frequency for various types of read and extemporaneous speech material. Qualitative examination of these data reveals spectra and distributions that are distinctive for an individual under specified conditions, but we are collecting further data that will indicate the extent to which these measures remain stable for a given individual over weeks and months. We plan to try to relate particular attributes of these statistical characteristics of a speaker to detailed properties of his anatomy or to his speaking habits, for example, the stability and position of higher formant frequencies, and the general form of the speaker's "breath group" intonation contour.