XXIV. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens
Prof. M. Halle
Prof. W. L. Henke
Prof. A. V. Oppenheim
Dr. Mary C. Bateson
Dr. Margaret Bullowa
Dr. A. W. F. Huggins
Dr. D. H. Klatt
Dr. Paula Menyuk
Dr. T. H. Shriner
Dr. J. J. Wolf

Graduate Students

R. E. Albright
Kay Atkinson
A. J. Goldberg
J. S. Perkell
J. E. Richards
R. M. Sachs

RESEARCH OBJECTIVES AND SUMMARY OF RESEARCH

The broad aim of our research in speech communication is to gain an understanding of the nature of the processes of human speech production and perception. A practical goal is to utilize knowledge gained through study of these processes to devise procedures that will permit limited communication between men and machines by means of speech. Several projects directed toward these goals are active at present. Studies of the relations between articulatory configurations and the speech sounds generated by these configurations, as well as experiments on the perception of speechlike sounds, are providing new evidence for the existence of discrete categories of speech sounds and are leading to further examination and modification of the system of distinctive features underlying speech events.[1-4] Studies of the production and perception of speech sounds by children, by using spectrographic techniques and tests involving the responses of children to natural and synthetic speech stimuli are providing some insights into the process whereby language is acquired.[5-7] Work is being carried out to develop computer-generated spectrographic displays and to implement the analysis and enhancement of vocal tract x-ray pictures using our computer facility. Other projects have been devoted to problems of automatic speech recognition[8] and speaker recognition.[9] A series of studies has been started to investigate the perceptual integration of signals switched alternately to the left and right ears.[10] Much of our research is now being carried out with the aid of a digital computer facility and associated peripheral equipment, including a graphical input, displays, a filter bank, a speech synthesizer, and other facilities.[11-13] We are continuing to improve and develop this system, and we expect in the forthcoming year to use the computer facility for further studies of the simulation of articulatory processes in speech perception, as well as for projects of the kind outlined here.

K. N. Stevens, M. Halle, A. V. Oppenheim, D. H. Klatt

References

1. M. Halle, and K. N. Stevens, "On the Feature 'Advanced Tongue Root'," Quarterly Progress Report No. 94, Research Laboratory of Electronics, M.I.T., July 15, 1969, pp. 209-215.
2. D. H. Klatt and K. N. Stevens, "Pharyngeal Consonants," Quarterly Progress Report No. 93, Research Laboratory of Electronics, M.I.T., April 15, 1969, pp. 207-216.

*This work was supported in part by the U.S. Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract F19628-69-C-0044; and in part by the National Institutes of Health (Grant 2 RO1 NB-04332-07).

3. R. M. Sachs, "Vowel Identification and Discrimination in Isolation and Word Context," Quarterly Progress Report No. 93, Research Laboratory of Electronics, M.I.T., April 15, 1969, pp. 220-229.

4. S. Cushing, "English as a Tone Language: The Acoustics of Primary Stress," Quarterly Progress Report No. 92, Research Laboratory of Electronics, M.I.T., January 15, 1969, pp. 351-359.

5. Paula Menyuk, Sentences Children Use (The M.I.T. Press, Cambridge, Mass., 1969).

6. Paula Menyuk and Suzan Anderson, "Children's Identification and Reproduction of /w/, /r/, and /l/," J. Speech Hearing Res. 12, 39-52 (1969).

7. Paula Menyuk, Acquisition and Development of Language, Text in a series entitled "Current Research in Child Development" (Prentice-Hall Inc., Englewood Cliffs, N.J., in press).

8. M. F. Medress, "Computer Recognition of Single-Syllable Words Spoken in Isolation," Quarterly Progress Report No. 92, Research Laboratory of Electronics, M.I.T., January 15, 1969, pp. 338-351.

9. J. J. Wolf, "Acoustic Measurements for Speaker Recognition," Quarterly Progress Report No. 94, Research Laboratory of Electronics, M.I.T., July 15, 1969, pp. 216-222.

10. A. W. F. Huggins, "Perceptual Integration of Dichotic Click Trains," Quarterly Progress Report No. 92, Research Laboratory of Electronics, M.I.T., January 15, 1969, pp. 359-362.

11. W. L. Henke, "Speech and Audio Computer-Aided Examination and Analysis Facility," Quarterly Progress Report No. 95, Research Laboratory of Electronics, M.I.T., October 15, 1969, pp. 69-73.

12. W. L. Henke, "TASS – Another Terminal Analog Speech Synthesis System," Quarterly Progress Report No. 95, Research Laboratory of Electronics, M.I.T., October 15, 1969, pp. 73-81.

13. A. W. F. Huggins, "A Facility for Studying Perception of Timing in Natural Speech," Quarterly Progress Report No. 95, Research Laboratory of Electronics, M.I.T., October 15, 1969, pp. 81-83.

A. MUTUALLY COMPLEMENTARY EFFECT OF RATE AND AMOUNT OF FORMANT TRANSITION IN DISTINGUISHING VOWEL, SEMIVOWEL, AND STOP CONSONANT

Previous research (Liberman[1]) has shown that the length of the time interval in which formant frequencies are changing is an important cue for distinguishing among vowel, semivowel, and stop consonant. The purpose of the present research is to determine whether the reported effect depends on rate of formant motion, amount of formant motion, or on both factors.

Synthesized wordlike sounds /aba/, /awa/, /aua/, and /aaa/ were used as the experimental materials because the motion of the first-formant frequency is the main cue for distinguishing among them, and accordingly, the first-formant frequency could be used as the only control factor in the experiment. Eight different transition patterns and

11 target values between 250 Hz and 790 Hz were prepared as x in a wordlike sound /axa/. The results indicate that an increase in the rate of first-formant frequency change reduces the amount of frequency change required to switch the identification to a particular consonant.

1. Procedure

In order to prepare the necessary experimental material, a series-connected terminal analog speech synthesizer was used (Tomlinson[2]). The synthesizer was controlled by a digital computer with graphic input/output capability (Henke[3]). Stimulus tapes were recorded and later played to subjects in a soundproof listening room. Figure XXIV-1 is a block diagram of the system.

In all of the experiments described in this report, a stimulus was an utterance of 500-ms duration that began and ended with the vowel [a], and was voiced throughout. The

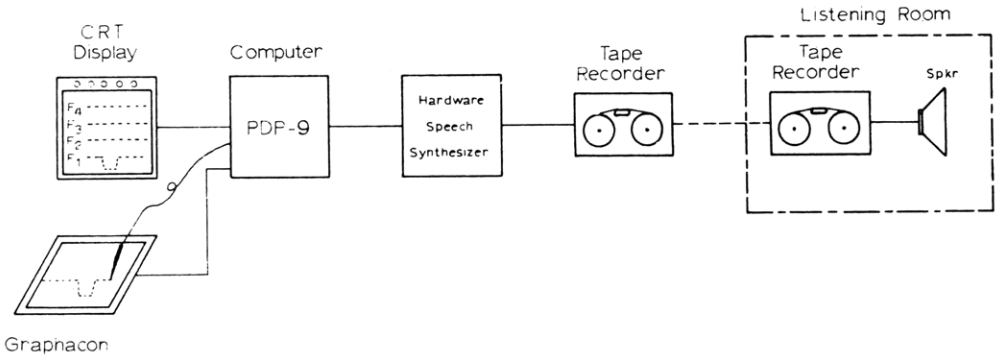

Fig. XXIV-1. Experimental apparatus.

only parameters that were varied in order to generate a sequence of stimuli was the location and the motion of the first-formant frequency during the utterance. The frequencies of the second, third, and fourth formants were always fixed at 1200 Hz, 2700 Hz, and 3300 Hz, respectively. The bandwidths of the first three formants were 73 Hz, 80 Hz, and 166 Hz. The contour of the fundamental frequency is shown in Fig. XXIV-2. A smoothly falling fundamental frequency was chosen to give the stimuli a natural intonation pattern that does not possess a constant fixed relationship between formant location and fundamental frequency.

Figure XXIV-3 illustrates the construction of a set of 11 stimuli for use in one of the experiments. $F1_k(t)$ is the control function for the first-formant frequency corresponding to stimulus number k. Stimuli are generated by interpolating between a limiting control function $F1_\phi(t)$ and a constant frequency, 790 Hz, by the formula

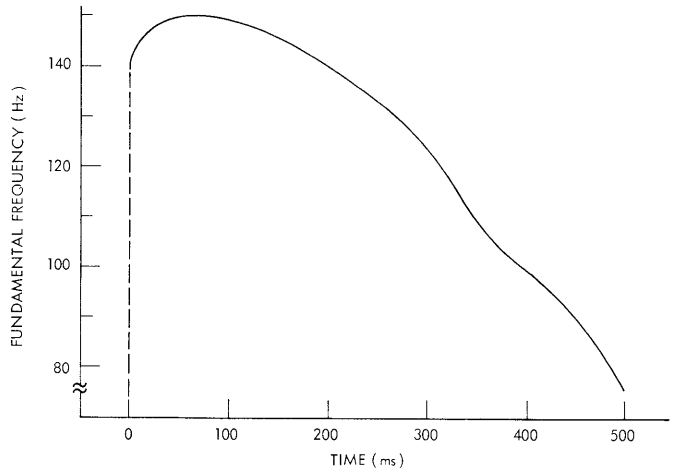

Fig. XXIV-2. Fundamental frequency contour used in the synthesis of stimuli for all experiments.

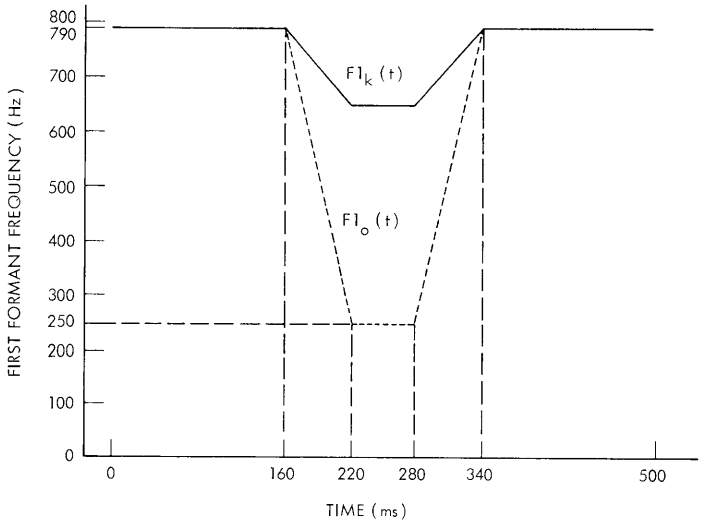

Fig. XXIV-3. Illustration of the procedure for constructing a time function to control the first-formant frequency. The function $F1_k(t)$, which controls the frequency of stimulus number k, is constructed by interpolation between a reference time function, $F1_o(t)$, corresponding to stimulus number k = 0, and the 790-Hz target frequency of the surrounding [a] vowel.

$$Fl_k(t) = \frac{10 - k}{10} Fl_\phi(t) + \frac{k}{10} \quad 790 \text{ Hz.} \tag{1}$$

A spectrogram of the synthesis for stimulus k = ϕ is shown in Fig. XXIV-4.

In each experiment, a randomized sequence of 22 test stimuli was presented. Eleven different sounds, each of which corresponded to a particular k in (1) appeared twice. Subjects were instructed to identify the sound segment placed in the middle of a stimulus pattern from an ensemble of responses /a, u, w, b/. (A few [m, r, ℓ] responses were given by some subjects when an open response set was used.) The pauses between stimuli were of 1-s duration. Subjects listened to a tape-recorded 22-item sequence 3 times in each experiment. A practice session preceded the first test. The listeners used in this experiment were 7 native speakers of Japanese, 5 male and 2 female. Therefore each stimulus pattern was heard 42 times.

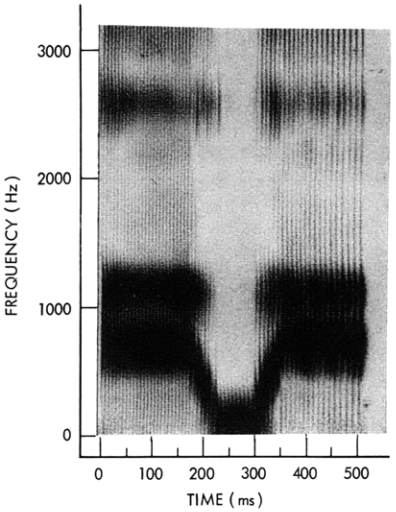

Fig. XXIV-4.
Spectrogram of the stimulus corresponding to k = 0 in the experiment described in the text and illustrated in Fig. XXIV-3.

The resulting identification function for a single subject (HS) is shown in Fig. XXIV-5. (This subject heard the recorded sequence 5 times.) Note that the boundary between

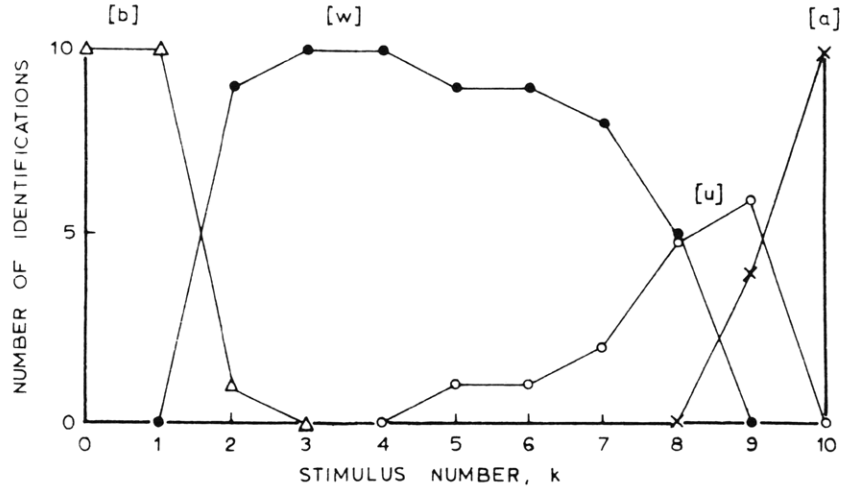

Fig. XXIV-5. Identification function of a single subject for the experiment shown in Fig. XXIV-3.

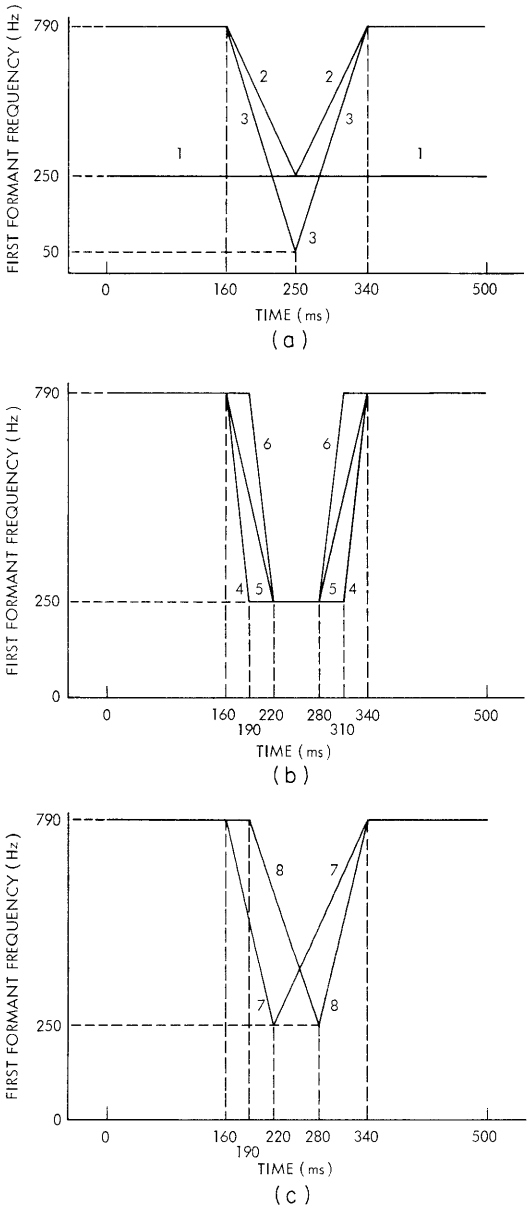

Fig. XXIV-6. Experiments 1-8 are defined by specifying a reference first-formant control function, $F1_{\phi}(t)$, for each.

predominantly /w/ identifications and predominantly /b/ identifications occurs at stimulus number 1.5. Referring to Fig. XXIV-3, this boundary would correspond to a stimulus having a rapid downward shift in first-formant frequency from 790 Hz to 331 Hz in 60 ms followed by a rapid upward shift in formant frequency from 331 Hz to 790 Hz in 60 ms. This kind of information will be used to construct equal-perception contours in a two-dimensional space involving rate of frequency change and absolute amount of frequency change.

In this study, 8 different transition patterns, $F1_{\phi}(t)$, were used to generate stimuli for 8 different experiments. In Fig. XXIV-6, the frequency of the first formant for signal k = 0 on each of the experiments is shown as a function of time. (Note that experiment number 5 was described in Fig. XXIV-3.) The set of first-formant reference patterns was chosen to include patterns with different values of frequency change as a function of time (Fig. XXIV-6a), different lengths of a central constant interval (Fig. XXIV-6b), and different nonsymmetrical increasing and decreasing formant frequency shifts (Fig. XXIV-6c). As can be seen from Fig. XXIV-6, the first experiment is a simple vowel categorization task.

2. Results

The results of the listening tests were plotted in terms of identification scores for each phoneme against the value of the stimulus parameter k (an example of this kind of plot is shown in Fig. XXIV-5). The average data of 7 subjects were combined because subjects displayed similar response patterns, and the data from any one subject were not sufficient to establish perceptual boundaries that are statistically significant. In every stimulus sequence (except, of course, the sequence in experiment 1), the identification scores shift systematically from /b/ to /w/ to /u/ to /a/ as the value of k becomes large.

To make clear the relationship between changing values of k and patterns of formant transition in each test sequence, all of the phoneme boundaries between stimuli are plotted in Fig. XXIV-7 as the points in a two-dimensional space involving ΔF_1 vs $\Delta \dot{F}_1$, where ΔF_1 is the absolute value of the first-formant frequency change, and $\Delta \dot{F}_1$ is the rate of formant change with respect to time. The values of k defining phoneme boundaries between /a/ and /u/, /u/ and /w/, and /w/ and /b/ taken from graphs of averaged data from 7 subjects, are indicated in this figure.

The representation of Fig. XXIV-7 does not distinguish an intial downward transition of the first-formant frequency from the later upward frequency transition; therefore, response boundaries in experiments 7 and 8 are plotted as two different points, one of which corresponds to the rising interval in the stimulus (indicated by an upward-directed arrow) and the other to the falling interval (indicated by a downward-directed arrow).

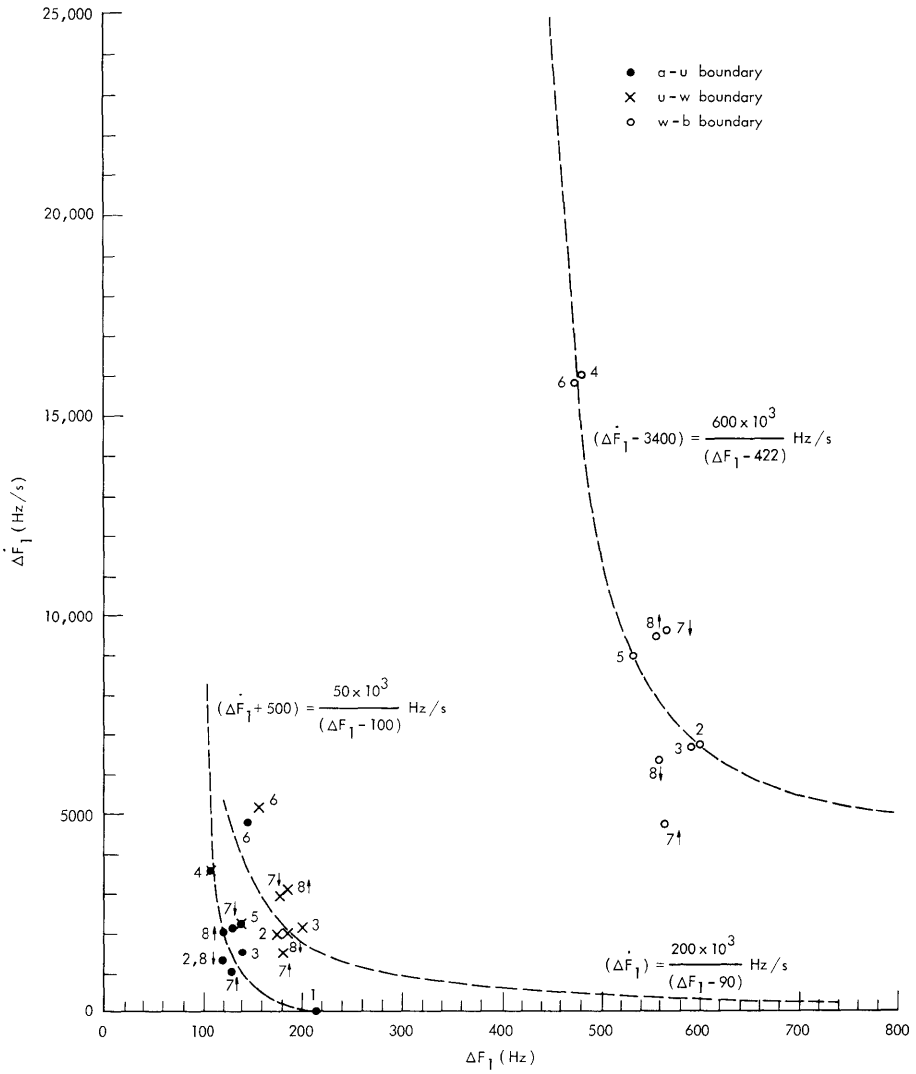

Fig. XXIV-7. Response boundaries between [a] and [u], [u] and [w], and [w] and [b] are plotted in terms of total absolute change in first-formant frequency, $|\Delta F_1|$, and the magnitude of the rate of frequency change, $|\Delta\dot{F}_1|$. The number adjacent to each point indicates the experiment from which the point is derived.

3. Discussion

a. Phoneme Boundaries

The dotted lines in Fig. XXIV-7 are hyperbolic curves that are proposed as the hypothetical phoneme boundaries. There are several reasons why hyperbolic curves are hypothesized. First, a phoneme boundary is likely to be a continuous monotonically decreasing line in the ΔF_1 vs $\Delta \dot{F}_1$ plane. Second, the experimentally obtained phoneme boundaries show a tendency that high ΔF_1 is accompanied by low $\Delta \dot{F}_1$; that is, there is an inverse relationship between ΔF_1 and $\Delta \dot{F}_1$. The third is derived from a subsidiary experiment in which the sounds having very high $\Delta \dot{F}_1$ and low ΔF_1, and very low $\Delta \dot{F}_1$ and high ΔF_1 were synthesized. If the ΔF_1 is under approximately 100 Hz, the sounds with even very high $\Delta \dot{F}_1$ were not identified as one of the ensemble /a, u, w, b/. On the other hand, the sounds having very low $\Delta \dot{F}_1$ and high ΔF_1 were identified not as /b/ but as /w/. This result implies that there is some threshold value for $\Delta \dot{F}_1$. The hyperbolic function

$$(\Delta \dot{F}_1 - A) = \frac{C}{(\Delta F_1 - B)}, \tag{2}$$

in spite of its simplicity, satisfies the requirements mentioned above. The constants A, B, and C were decided from the data as follows:

Phoneme boundary between /b/ and /w/: $(\Delta \dot{F}_1 - 3400) = \frac{(600 \times 10^3)}{(\Delta F_1 - 422)}$ Hz/s,

Phoneme boundary between /w/ and /u/: $(\Delta \dot{F}_1) = \frac{(200 \times 10^3)}{(\Delta F_1 - 90)}$ Hz/s,

Phoneme boundary between /u/ and /a/: $(\Delta \dot{F}_1 + 500) = \frac{(50 \times 10^3)}{(\Delta F_1 - 100)}$ Hz/s.

b. Effect of Changing the Central Constant Interval

The reference patterns of first-formant change in experiments 4 and 6 have the identical rate of frequency change, but different lengths of a central constant interval. Comparing the value of k for phoneme boundaries in those experiments, it is found that the shorter the central constant interval, the larger the frequency change at the boundaries between /a/ and /u/, and /u/ and /w/. However, the boundary between /b/ and /w/ is not influenced by the length of central constant interval.

The reference pattern of first-formant change in experiments 2, 4, and 5 have different rates of change, as well as different lengths of a central constant interval. It has been stated that, in general, the higher the rate of frequency change, the lower the amount of frequency change in the phoneme boundary. As far as these experiments are concerned, however, this relationship is prominent in the boundary between /b/ and /w/, and is less prominent in the boundaries between /w/ and /u/, and /u/ and /a/.

From these findings, it can be said that the perception of a stop consonant is influenced by the rate of frequency change and is relatively insensitive to changes in the length of the central constant interval, whereas the perception of semivowel and vowel is influenced largely by the length of the central constant interval, as well as the rate of frequency change.

c. Effect Resulting from the Difference between Downward Transition and Upward Transition

Among the 8 experiments in this study, only experiments 7 and 8 have nonsymmetrical increasing and decreasing formant frequency shifts. The fact that the curve for phoneme boundary between /b/ and /w/, which is more sensitive than other phoneme boundaries, passes through between 7↓ and 7↑, and 8↑ and 8↓ in Fig. XXIV-7 suggests that if the rate of the upward formant motion is different from that of downward, a virtual rate of formant motion is to be expected at some value between the upward and the downward rates.

The value of k for the phoneme boundary between /b/ and /w/ is smaller in experiment 7 (k = 1.1) than in experiment 8 (k = 1.25). This suggests that the upward transition has larger influence in the perception of stop consonant /b/ than the downward transition.

H. Suzuki

References

1. A. M. Liberman, P. C. Delattre, L. J. Gerstman, and F. S. Cooper, "Tempo of Frequency Changes as a Cue for Distinguishing Classes of Speech Sounds," J. Exptl. Psychol., Vol. 52, No. 2, August 1956.

2. R. S. Tomlinson, "SPASS – An Improved Terminal Analog Speech Synthesizer," Quarterly Progress Report No. 80, Research Laboratory of Electronics, M.I.T., January 15, 1966, pp. 198-205.

3. W. L. Henke, "Speech Computer Facility," Quarterly Progress Report No. 90, Research Laboratory of Electronics, M.I.T., July 15, 1968, pp. 217-219.