ANN.Bull.RILP(1990)No.24, 97-100

PERCEPTION OF FREQUENCY-COMPRESSED VOICED STOPS AND SEMIVOWELS

Sotaro Sekimoto

1. Introduction

The characteristics of the perceptual normalization for frequency-compressed speech have been studied. For vowels, it has been found that speech in which the frequency axis is compressed or expanded is identified as natural over a wide range of frequency expansion or compression ratios[1]. It has been concluded that a similar perceptual normalization, observed for ordinary male, female and children's voices, plays an important role in identifying frequency compressed speech. Although it is supposed that the characteristics of perceptual normalization again play a role in the perception of frequency-compressed consonants, few studies have been made on perceptual normalization with the compression of the frequency axis[2]. In the present study, the characteristics of perceptual normalization with the compression of a frequency axis in word-initial voiced stop consonants was examined.

It is well known that the perception of synthetic consonant-vowel syllables varies from /ba/ to /wa/ with changes in the extent, speed or duration of the transition of the lowest or two lowest formant frequencies[3][4][5]. On the other hand, the extent and speed of the formant-frequency transition becomes smaller if frequency-compression is made. Therefore, it is suspected that the perception of /ba/ sometimes changes to /wa/ when frequency-compression is made. In the present study, a perceptual experiment was conducted to examine whether such a change in the perception of /ba/ occurs.

2. Method

Synthetic speech stimuli in which the extent and the time-constants of the formant transitions were varied were used. The formant transition pattern was approximated by a step-response of the 1st-order linear system, since it was hard to recognize a stimulus as a stop consonant if the transition was approximated by a piece-wise linear curve. The frequency (f_n) at t was defined as

$$f_n(t) = F_{ne} - F_{ns} \exp(-t/T_c)$$

where F_{ne}, F_{ns}, T_c were the target frequency of the following vowel, the extent of frequency transition and the time-constant of a transition, respectively.

The experimental parameters were as follows.

(1) The time-constants of the first and the second formant transitions: from 2 msec to 30 msec in 2-msec steps (15 steps in all).

(2) The extent of the first formant transition: 300, 400, 500, 600 and 700 Hz.
(3) The extent of the second formant transition: 0 and 400 Hz. In the case of 400 Hz, the first and the second formant frequencies were varied concurrently for the same time-constant.
(4) The frequency-compression ratio, defined as a percentage of the uncompressed condition (100%) : 100, 80 and 60 %.

Examples of the formant transition patterns are shown in **Fig. 1**. The vowel was /a/.

The stimuli were synthesized on the software terminal-analog speech synthesizer shown in **Fig. 2**. Rosenberg's C-waveform was used as a voice source[6]. The target frequencies of the formant transitions, or the formant frequencies, of the following vowel /a/ in the uncompressed condition were 720, 1240, 2500, 3500 and 4500 Hz, respectively for F1, F2, F3, F4 and F5. The bandwidth of each formant was 50, 70, 110, 200 and 250 Hz, respectively.

The duration of each stimulus was 300 msec, with a 10-msec leading increase and trailing decrease in amplitude across 80dB. The fundamental frequency was kept constant at 110 Hz for each frequency-compression ratio.

The compression of the frequency axis was accomplished by lowering the sampling frequency of the filters of the synthesizer. The sampling frequency for the uncompressed condition was 10 kHz. The speech stimuli were produced on a software synthesizer and D/A-converted at 12-bit precision. The signal was then low-pass filtered with a cutoff of -135 dB/oct. and recorded on a DAT (Digital Audio Tape). The cutoff frequency of the low-pass filter was dynamically changed in proportion to the sampling frequency by a factor of 0.45.

The stimuli were presented through binaural headphones (STAX SR-Lambda Signature) in a soundproof room at a comfortable presentation level (about 75 dB SPL).

The speech stimuli were presented in a random order. Subjects were requested to identify the synthetic stimuli as one of the Japanese vowels, voiced stops or semivowels (including /wo/). Three adult subjects participated.

3. Results and discussion

The perceptual phonetic boundary between vowels, stops and semivowels when the extent of the second formant transition was 400 Hz are shown in **Fig. 3**. The ordinate shows the extent of the first formant transition before the frequency-compression was made. The perceived phoneme varied from /a/ to /ba/ to /wa/ as the time-constant increased when the frequency compression ratio was 100 and 80 %, and it varied from /o/ to /bo/ to /wo/ when the frequency compression ratio was 60%. Thus, as the frequency

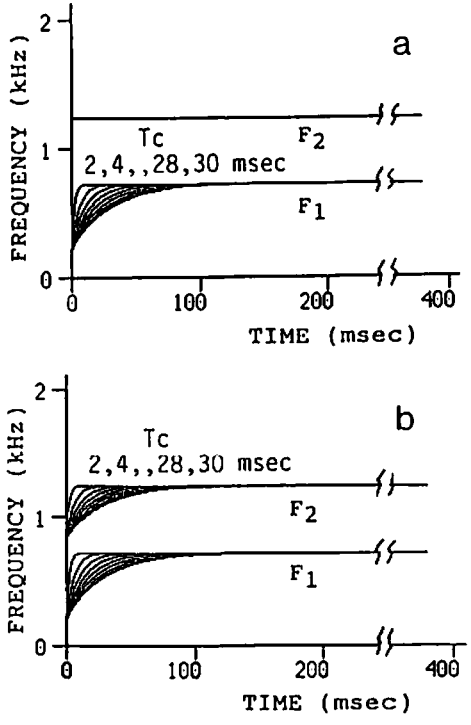

Fig. 1. Examples of the first and second formant transition patterns for various time-constants in synthesizing /b-w/. Fig. a shows the patterns when F2 was constant, and Fig. b shows those when F1 and F2 were varied concurrently.

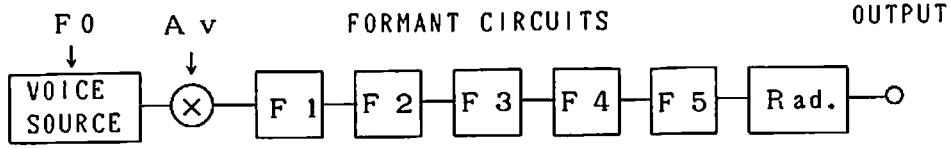

Fig. 2. A block diagram of the software speech synthesizer.

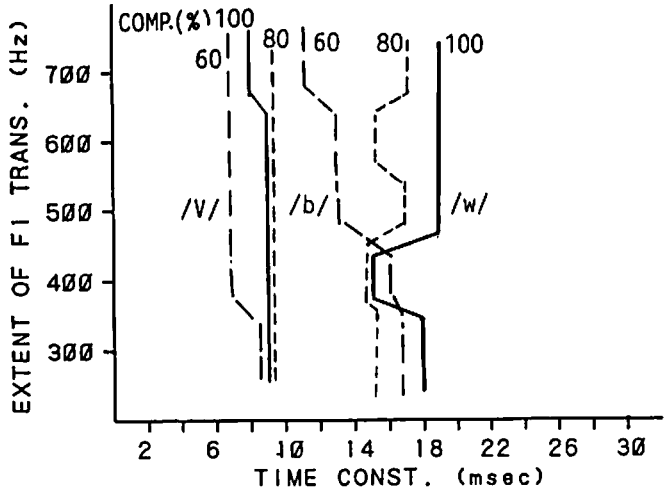

Fig. 3. Perceptual phoneme boundaries on the plane of the F1-change vs. the time-constant when F1 and f2 were varied concurrently with the same time-constant. The phoneme boundaries were defined as 50% identification contours.

compression ratio decreased, the /b-w/ boundary shifted toward time-constants of smaller value in the condition where the extent of the first formant transition was above 400 Hz. This result agrees with our initial expectation.

When the extent of the second formant frequency was zero, an identification as /ga/ appeared for the frequency compression ratio of 80%, so this result is omitted. It seems that the perception of stop consonants was not stable if F2 was constant, and that the perceptual boundary of the second formant frequency was shifted downward by the frequency compression, causings the identification to change. This result suggests that at least the two lowest formant frequencies should be changed in order to maintain place of articulation in the perception of frequency-compressed voiced stops.

References

1) Sekimoto, S. (1982); Perceptual normalization of frequency scale, Ann. Bull. RILP, 16, 95-101.
2) Sekimoto, S. (1989); Normalization of frequency-compressed voiceless fricatives, Ann. Bull. RILP, 23, 39-49.
3) Liberman, A. M., P. C. Delattre, L. J. Gerstman, and F. S. Cooper (1956); Tempo of frequency change as a cue for distinguishing classes of speech sounds, J. Exp. Psychol., 52, 127-137.
4) Suzuki, H. (1974); Mutually complementary effect between amount and rate of formant transition in perception of vowels, semivowels and voiced stops and a possible mechanism for their identification, J. Acoust. Soc. Japan, 30, 169-180.
5) Schwab, E. C., J. R. Sawusch, and H. C. Nusbaum (1981); The role of second formant transitions in the stop-semivowel distinction, Psychophysics, 29, 121-128.
6) Rosenberg, A. E. (1971); Effect of glottal pulse shape on the quality of natural vowels, J. Acoust. Soc. Am., 49, 583-590.