ANN.BULL.RILP(1991)No.25. 103-108

The role of gross spectral shape as an invariant cue for identifying voiceless stop consonants

Sotaro Sekimoto

1. Introduction

The characteristics of perceptual normalization for the compression or expansion of the frequency axis in the identification of the place of articulation of voiceless stop consonants have been studied.[1] In these studies, the contribution of two properties in the pre-vocalic noise of voiceless stop consonants to normalization have been explored. One property has been the number of spectral components that approximate the noise structure. Two conditions for the number of spectral components have been examined: the single-pole condition and the multiple-pole condition, where the noise structure is approximated by single-pole and multiple-pole noises, respectively. These two conditions have been adopted to determine whether normalization is made absolutely from the noise center frequency or relatively among the noise formant components. Perceptual experiments to identify the place of articulation of voiceless stop consonants have been carried out using synthetic stimuli in which the frequency axes have been compressed or expanded. Results have shown that the perceptual phoneme boundary between /k/ and /t/ shifts along with the variation in the frequency-compression or expansion ratio in both conditions. These results suggest that frequency normalization is made in the perception of voiceless stops. The difference between the two conditions appears merely in the inclination of the shift against the frequency-compression or expansion ratio, that is, the inclination is steeper for multiple-pole noise than for single-pole noise.

From such results, it may seem that the difference in the number of spectral components does not significantly affect normalization. However, the question whether normalization is made absolutely or relatively from the spectral components of noise is still unresolved.

Another property which has been studied is the frequency transition of noise components. Synthetic speech stimuli with or without frequency transitions has been subjected to hearing tests. A contribution of frequency transitions to the perceptual normalization of the frequency axis has not been observed.

From these experiments, it has been found that neither the number of spectral components nor a frequency transition in pre-vocalic noise is an essential cue of the normalization for the compression or expansion of the frequency axis. However, from the results of the first experiment, it is reasonable to expect that the gross spectral shape of the pre-vocalic noise of voiceless stop consonants is one cue towards identifying voiceless stop consonants because the gross spectral shape is similar for single and multiple poles. The difference in the number of spectral components seems to act as a difference of the "weight" on the gross spectral shape. Actually, it has been reported from acoustic analyses of natural speech and from perception experiments that the invariant property

for the identification of the place of articulation of initial stop consonants is the gross shape of the spectrum in the first 20-odd ms of the signal[2,3].

If the cue is invariant, a normalization for the difference in the frequency axis is not needed and the same experimental result will be obtained because the cue detection and the normalization are accomplished simultaneously in the same process. Therefore, in the present study, the contribution of the gross spectral shape of the pre-vocalic noise of voiceless stop consonants as a invariant cue was examined. An identification test for voiceless stop consonants was carried out using frequency-compressed or expanded synthetic speech stimuli. Here, the difference in the frequency-compression or expansion rate acted as a source of variance in the vowel environment. Using frequency-compressed or expanded synthetic speech, the context effect which caused by the vowel identification itself could be avoided. The gross spectral shape of the pre-vocalic noise was simulated by a low-pass or high-pass filtered noise.

2. Method

The identification rates for the Japanese voiceless stop consonant /p/, /t/ and /k/ were measured for combinations of the following experimental parameters.

Experimental parameters

(1) Cutoff frequencies of the filter (Fc)
(a) Lowpass filter:
800, 1000, 1200, 1400, 1600, 1800, 2000, 2200, and 2400 Hz.
(b) Highpass filter:
800, 1200, 1600, 2000, 2400, 2800, 3200, 3400, and 4000 Hz.

(2) Frequency-compression or expansion ratios (defined as percentage of the frequency-compression or expansion of the natural condition)
60% (maximally compressed), 80% (compressed), 100% (uncompressed), 120% (expanded), and 140% (maximally expanded).

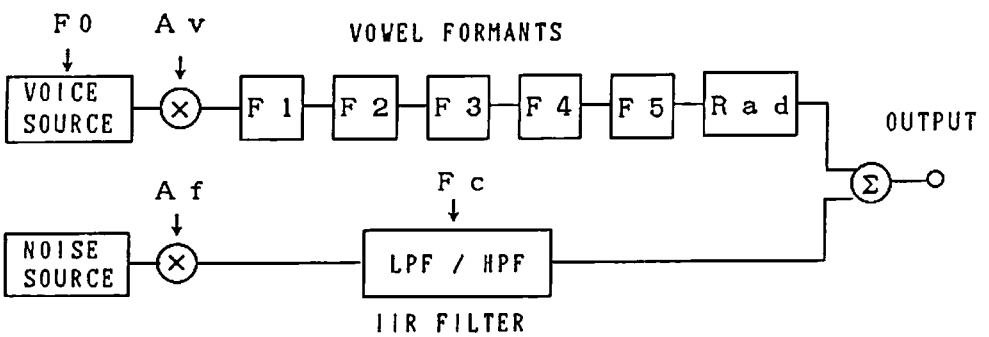

Fig. 1. A block diagram of the software speech synthesizer.

Stimuli

Stimuli were synthesized with a software terminal-analog speech synthesizer. A block diagram of the synthesizer is shown in **Fig. 1**. The noise period for the voiceless stops was synthesized using a lowpass (LPF) or a highpass (HPF) filter in the lower branch. These filters were sixth-order IIR filters with Butterworth characteristics, and were realized as cascaded constructions of three 2nd-order filters. The resulting slope of the cut-off was -36 dB/oct. The frequency responses of the lowpass and highpass filters for each cutoff frequency are shown in **Fig. 2**. Rosenberg's C-waveform[4] was used as a voice source. The temporal control patterns of the synthesizer for the lowpass condition are shown in **Fig. 3** as an example. The nine conditions of the cutoff frequency of the lowpass filter (Fc) are drawn in dotted lines. The initial 10ms period simulated the initial

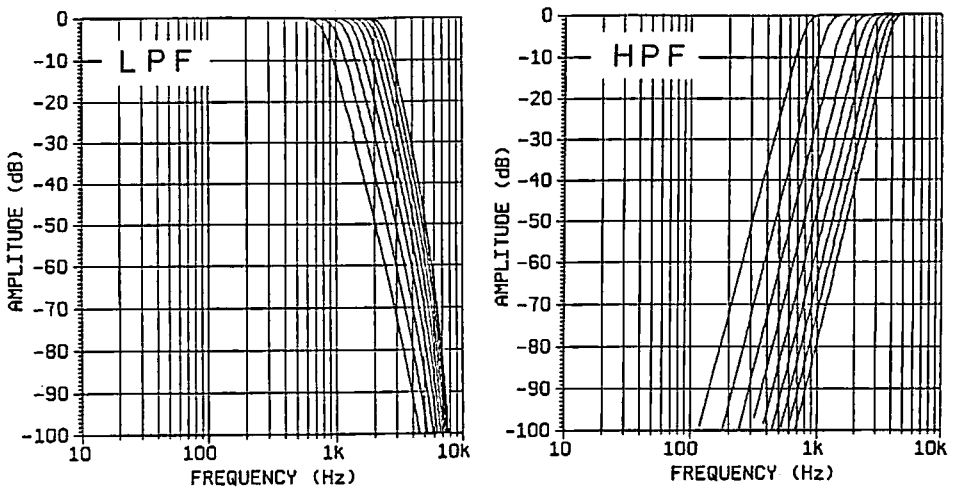

Fig. 2. The frequency responses of lowpass and highpass filters which simulate the initial noise period of synthetic voiceless stop consonants.

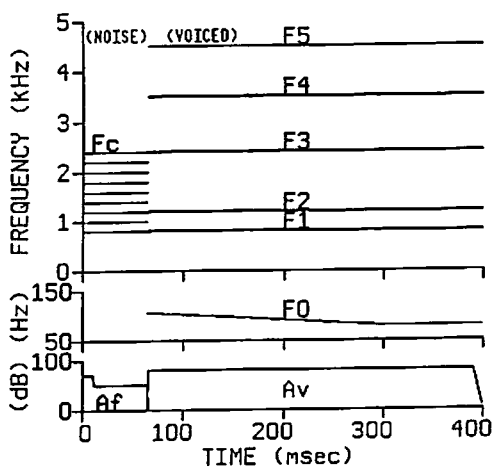

Fig. 3. The time patterns of the control parameters for synthesizing /ta/-/ka/.

noise burst of voiceless stops. The next 55ms was a noise period which simulated aspiration. The RMS amplitude of the aspiration period was set so as to be 20 dB less than that of the burst period. The vocalic period then followed these noise periods. The vowel was always /a/. The formant frequencies of the vocalic portion, F_1, F_2, F_3, F_4, and F_5, were held constant at 800Hz, 1200Hz, 2400Hz, 3500Hz, and 4500Hz, respectively. The fundamental frequency was changed from 114Hz to 80Hz in the initial 300 ms period, then held constant until 400ms. The absolute fundamental frequency was independent of the frequency-compression or expansion ratio. The amplitude of the vocalic period was set so as to be 17 dB higher for the lowpass condition, and 26 dB higher for the highpass condition, than that of the initial burst period.

The compression or expansion of the frequency axis was accomplished by raising or lowering the sampling frequency of the synthesizer. The sampling frequency when the frequency axis was neither compressed nor expanded was 20 kHz. The synthetic digital speech wave was quantized in 12 bits and was converted to an analog wave. The analog wave was then low-pass filtered with a cutoff of -135 dB/oct., to eliminate the undesirable high frequency components caused by the frequency aliasing effect, 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.

Procedure

The stimuli were presented in random order through binaural headphones (STAX SR-Lambda Signature) in a soundproof room at a comfortable presentation level (about 75 dB SPL). Subjects were requested to identify the stimuli as one of the Japanese voiceless stops, /t/, /k/ or /p/.

3. Results and discussion

The results when the lowpass-filtered noise was used to simulate the initial noise burst and the aspiration noise are shown in **Fig. 4**. The phoneme which was identified most often for each combination of the cutoff frequency and the frequency-compression or expansion ratio is shown. The ordinate shows the absolute cutoff frequency of the filter after the frequency-compression or expansion ratio is multiplied. The perceptual phoneme boundary between /t/ and /k/, along with the cutoff frequency, varied with the frequency-compression or expansion ratio. The perceptual boundary between /k/ and /p/, on the other hand, showed an almost constant cutoff frequency, despite the variation in the frequency-compression or expansion ratio.

The results when the highpass-filtered noise was used to simulate the initial noise burst and the aspiration noise are shown in **Fig. 5**. The phoneme identification hardly changed with respect to the variation in the frequency-compression or expansion ratio. The perceived phoneme was always /t/ when the frequency-compression or expansion ratio was below 120%.

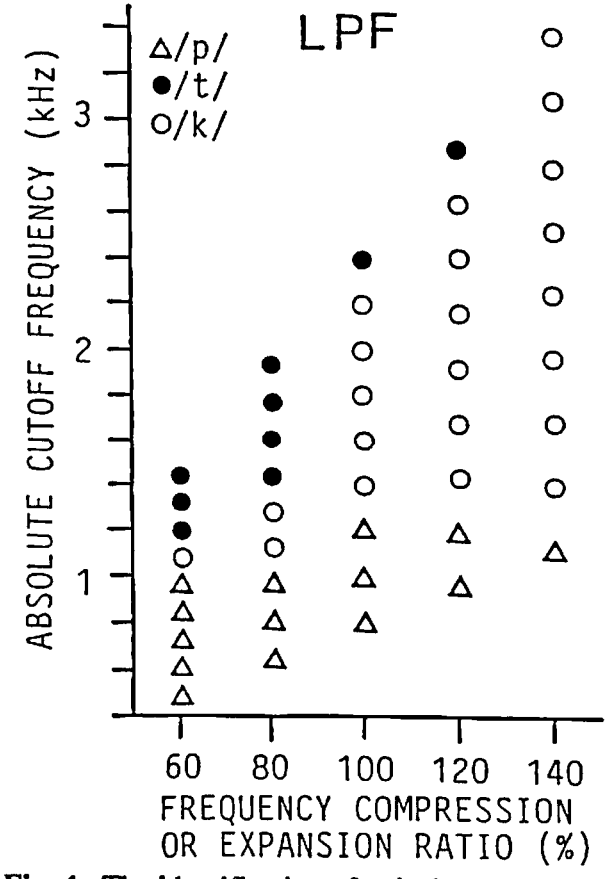

Fig. 4. The identification of voiceless stops when the initial noise burst and aspiration were simulated by a lowpass-filtered noise.

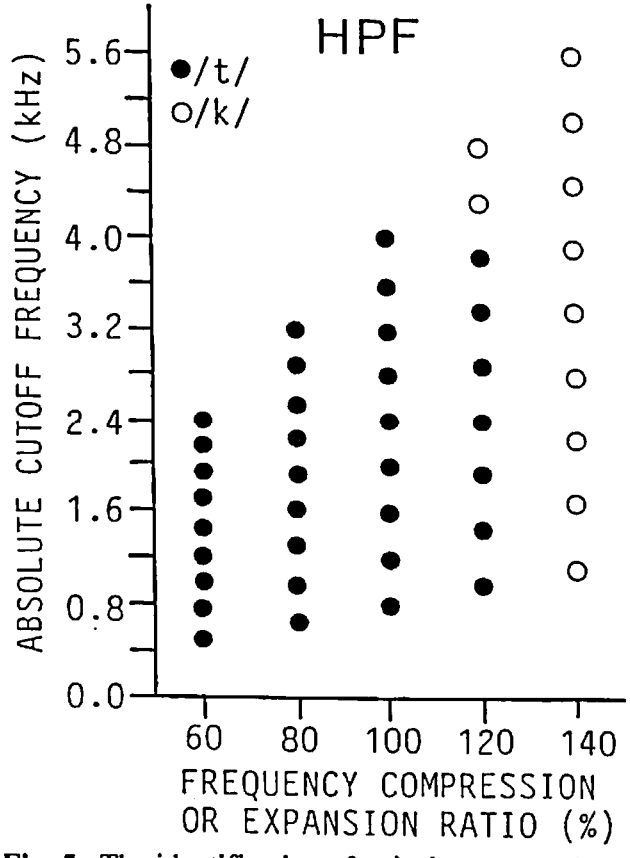

Fig. 5. The identification of voiceless stops when the initial noise burst and aspiration were simulated by a highpass-filtered noise.

These results suggest the existence of some property detector which derives a /t/ decision if the components of the initial pre-vocalic noise are distributed above a particular frequency, and a /p/ decision if the components of the initial pre-vocalic noise are distributed below a particular frequency. In natural speech, however, the inclination or declination of the gross spectral shape is not so sharp[3] as that used in this experiment. It can be considered that some section of the property detecting scheme for natural speech is reflected in our results. Further study is required to clarify the relationships between the data from this experiment and that from Stevens and Blumstein's work[2,3].

4. Conclusion

The results of this experiment suggest that the gross spectral shape of the pre-vocalic noise of voiceless stop consonants functions as a invariant cue in the identification of place of articulation.

References

1) Sekimoto, S. (1990); Perceptual frequency normalization of frequency compressed or expanded voiceless consonants, Proc. ICSLP 90, 589-592.
2) Stevens, K. N., and Blumstein, S. E. (1978); Invariant cues for place of articulation in stop consonants, J. Acoust. Soc. Am. 64, 1358-1368.
3) Blumstein, S. E., and Stevens, K. N. (1979); Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants, J. Acoust. Soc. Am. 66, 1001-1017.
4) Rosenberg, A. E. (1971); Effect of glottal pulse shape on the quality of natural vowels, JASA, 49, 2(Pt. 2), 583-590.