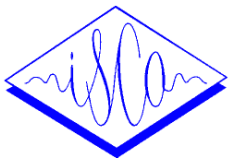

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PERCEPTUAL FREQUENCY NORMALIZATION OF FREQUENCY COMPRESSED OR EXPANDED VOICELESS CONSONANTS

Sotaro Sekimoto

Research Institute of Logopedics and Phoniatrics,
Faculty of Medicine, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113 Japan

ABSTRACT

In order to clarify the characteristics of frequency normalization on voiceless consonants in which the frequency axes were compressed or expanded, perceptual experiments were carried out for voiceless fricatives and voiceless stops. For voiceless fricatives, a series of fricative noises that varied from the center frequency appropriate for /ʃ/ to one appropriate for /s/ were synthesized as stimuli. These stimuli were subjected to a listening test to determine the phonetic categorical boundary between /ʃa/ and /sa/ on the continuum of the noise frequency for various frequency-compression or expansion rates. For voiceless stops, noise part was simulated by either a single-pole noise or a multiple-pole noise. A series of noises that varied from /k/ to /t/ were synthesized and subjected to a identification test between /ka/ and /ta/. The results showed that the categorical boundaries between /ʃa/ and /sa/ were invariant, but the categorical boundaries between /ka/ and /ta/ moved in relation to the frequency-compression or expansion rate for both noise conditions. These results implied that the frequency normalization was made on voiceless stops but not on voiceless fricatives.

1 INTRODUCTION

The efficiency of hearing aids with frequency-lowering has been investigated by the present authors[1]. The result of a hearing test, where the frequency axis of speech was compressed downward by a PARCOR speech analysis-synthesis method and the characteristics of a hearing impairment were simulated by a low-pass filter, showed that considerable improvement was observed for vowels, whereas little improvement was seen for consonants, and especially for voiceless consonants. It is not apparent, however, why such an improvement was or was not obtained. For vowels, it was observed in a subsequent study by the author that the speech of which the frequency axis was compressed or expanded was identified as original over a wide frequency expansion or compression ratio, especially when the fundamental frequency was concurrently raised or lowered in the same ratio[2]. Namely, the difference of the frequency axis was compensated for in the perceptual process, and perceptual frequency normalization occurred. These results suggest that perceptual frequency normalization plays an important role in identifying frequency compressed speech correctly. For consonants, on the other hand, although it is supposed that the characteristics of perceptual normalization again play a role in the perception of frequency-compressed consonant, few studies have been made on the normalization of consonant speech[3][4].

In the present study, perceptual experiments were performed to elucidate the characteristics of frequency normalization on frequency-compressed voiceless fricative and stops.

2 EXPERIMENT 1

2.1 Method

It is known that voiceless fricative consonants can be synthesized from a single-pole noise followed by a vowel portion[5]. In this case, the noise portion does not have a particular structure, such as a formant structure which is seen in vowel speech. Accordingly, if perceptual normalization occurs in the perception of voiceless fricatives, it can be assumed that the cue for the frequency normalization is not present in the noise portion itself but is in the relation between the noise portion and the following vowel portion.

Synthetic speech samples in which a single-pole frication noise portion was followed by a vowel portion were adopted to the hearing test to determine the perceptual boundary between /s/ and /ʃ/ on the continuum of resonant frequencies of the noise pole. The boundaries were compared for various ratios of the frequency compression.

Stimuli Stimuli were synthesized with a software terminal analog speech synthesizer. A block diagram of the synthesizer is shown in **Fig. 1**. Rosenberg's C-waveform was used as a voice source[6]. The time patterns of the control parameters of the synthesizer when the vowels /a/ and /u/ followed the noise are shown in **Figs. 2** (a) and (b), respectively. The formant pattern was close to that used in the experiment by Mann and Repp[7]. The same fundamental frequency patterns were used for both /a/ and /u/. The output level was set so that the level of the stationary portion of the noise was 12dB lower than that of the vowel for each stimulus. The compression of the frequency axis was accomplished by lowering the sampling frequency of the synthesizer filters. As a result, the spectrum envelope was compressed toward zero analogously. The sampling frequency without frequency compression was 10 kHz. The speech waveform synthesized on the Apollo DN-4000 workstation was D/A-converted at

12-bit precision and low-pass filtered with a cutoff of -135 dB/oct, and was 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.

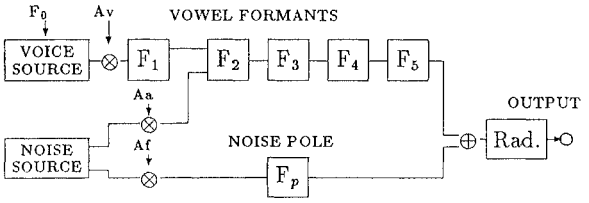

Fig. 1. A block diagram of the speech synthesizer.

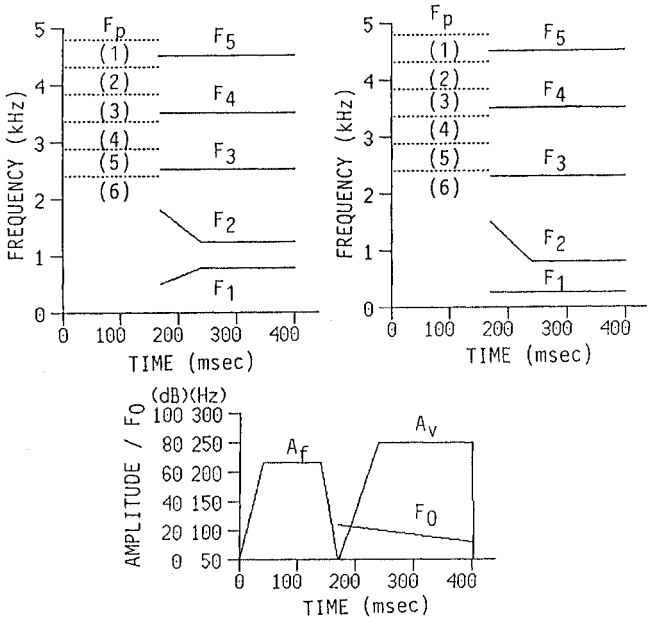

Fig. 2. The time pattern of the control parameters for synthesizing /sa/-/ʃa/ and /su-/ʃu/.

Experimental Conditions

- *The following vowel*: /a/ and /u/.
- *The center frequencies of noise*:
  2400, 2880, 3360, 3840, 4320, and 4800 Hz.
- *Frequency-compression ratio, which were defined as percent ratios of the frequency compression against the uncompressed condition*:
  100% (uncompressed), and 80% and 60% (compressed).

Procedure The speech material was presented through binaural headphones (STAX SR-Λ Signature) in a soundproof room. The presentation level was about 75 dB SPL. The speech samples were presented in a random order. Subjects were requested to identify the synthetic stimuli as one of the following Japanese syllables: /sa/, /su/, /so/, /ʃa/, /ʃu/, /ʃo/, /ha/, /fu/, /ho/. Each token was presented 20 times for each subject. Six adult subjects participated.

2.2 Results and discussion

The identification rates when the noise portion was followed by /a/ and /u/ are shown in **Figs. 3** (a) and (b), respectively. The answers of six subjects are averaged. The abscissa shows the absolute noise pole frequency after frequency compression was made. The ordinate shows the identification rates. The results for the three frequency compression ratios 60%, 80% and 100% (uncompressed condition) are shown in the same figure for the sake of comparison. The identification rates for /s/ and /ʃ/ are shown by solid lines and dotted lines, respectively.

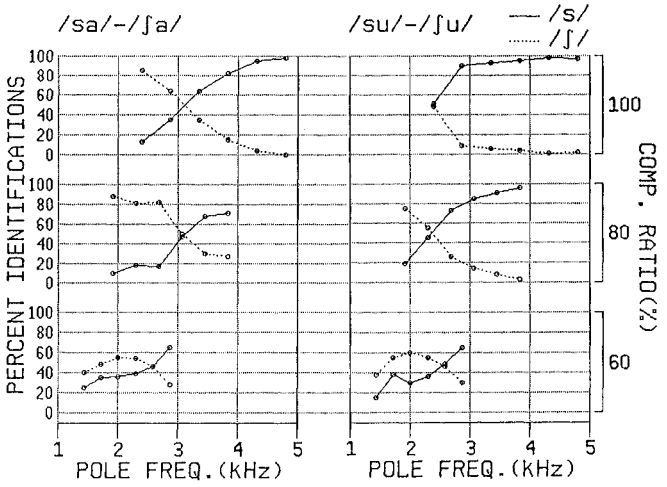

Fig. 3. The identification rates when the noise portion was followed by /a/ (left) and /u/ (right).

When the frequency-compression ratios were 100% and 80%, the boundary pole frequencies, where the responses of /s/ and /ʃ/ cross, were almost identical for both /a/ and /u/; however, their absolute boundary pole frequencies were different. The boundary frequency when the noise portion was followed by /u/ shifted downward compared with /a/. This result suggests the existence of a context effect from the following vowel on the identification of the prevocalic voiceless fricative consonant. This result agrees with that of Kunisaki and Fujisaki[5].

In the case where the frequency-compression ratio was 60%, the maximum identification rates for /s/ and /ʃ/ became lower because the response for /f/ increased. When the noise portion was followed by /a/, the boundary pole frequency between /s/ and /ʃ/ for the compression ratio 60% was lower than that for the compression ratios 80% and 100%. On the other hand, when the noise portion was followed by /u/, the boundary noise pole frequency was similar among the three frequency-compression ratios. Note that the following vowel was identified as /o/ for both vowel context conditions when the frequency compression ratio was 60%. In Kunisaki and Fujisaki[5], the perceptual boundary of the

noise pole frequency between /s/ and /ʃ/ was affected by the following vowel and the extent of the frequency shift was similar between /u/ and /o/ and between /a/ and /e/, but the extent was considerably different between the two groups. Thus, it can be concluded that the downward boundary shift when the noise portion was followed by /a/, which was exclusively observed for the frequency-compression ratio of 60%, is explained by the context effect of the following perceived vowel /o/.

These results suggest that the perceptual boundary is hardly changed by the frequency-compression as long as the identification of the vowel does not change, namely, the identification for voiceless fricative consonants is performed based on the absolute noise pole frequency.

Fig. 4 shows the supplementary result of an informal experiment where the frequency axis was compressed and expanded from 50% to 140% with fundamental frequency lowering and raising at the same ratios. It seems that the boundary pole frequencies between /s/ and /ʃ/ were consistent for a wide range of frequency-compression ratios. These data support the assumption that voiceless fricative consonants are identified from the absolute noise pole frequency.

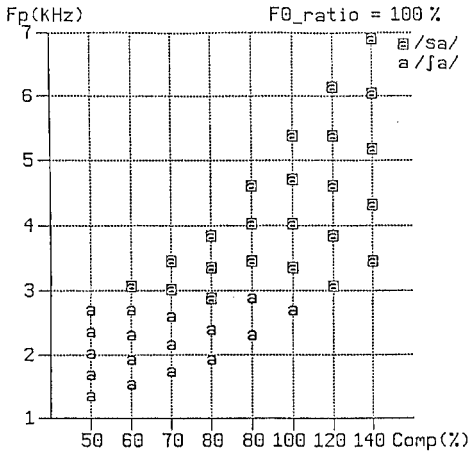

Fig. 4. Identification of /s/ and /ʃ/ at various frequency compression or expansion ratios.

3 EXPERIMENT 2

3.1 Method

In this experiment, two types of noise with different noise component structures were used. From a preliminary experiment, it was determined that the above two kinds of noise, that is, single-pole noise and multiple-pole noise, could be adopted as the prevocalic noise for a word-initial synthetic voiceless stop consonant. The single-pole noise was synthesized using the lower branch of the diagram shown in **Fig. 1** with attenuator "Af" switched to "On". The multiple-pole noise was synthesized using the upper branch where the attenuator "Av" was "Off" and "Aa" was "On". The former was the condition in which absolute identification was expected, and the latter was that where relative identification was expected.

The onset frequency of the resonant circuit, F_p for single-pole noise and F_2 for multiple-pole noise, was systematically varied and subjected to a hearing test to determine the perceptual phoneme boundary between /t/ and /k/. The phoneme boundaries were compared for various ratios of frequency compression or expansion. The vowel was /a/.

Stimuli Stimuli were synthesized with a software terminal analog speech synthesizer. Rosenberg's C-waveform was used as a voice source[6]. The time patterns of the resonant frequencies of the noise poles are shown in **Fig. 5**. The initial 10msec of the noise period was kept stable to simulate a noise burst, and the following 55msec was a transition period simulating aspiration. The transition pattern was approximated by a step-response of a 1st-order linear system. 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 the frequency transition and the time-constant of the transition, respectively. The time-constant was always 10msec. The value of the second formant frequency of the vocalic portion was used as a target. The values of the formant frequencies of the vocalic portion F_1, F_2, F_3, F_4, and F_5 were held constant at 800, 1200, 2400, 3500, and 4500Hz, respectively. The fundamental frequency was changed from 114Hz to 80Hz in the initial 300 msec period, then held constant until 400msec. The same fundamental frequency pattern was used for each frequency-compression or expansion ratio.

The compression or expansion of the frequency axis was accomplished by lowering or raising the sampling frequency of the filters of the synthesizer. The sampling frequency for the uncompressed condition was 20 kHz. The speech stimuli were synthesized 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. The cutoff frequency of the low-pass filter was dynamically changed in proportion to the sampling frequency by a factor of 0.45.

Experimental Conditions

- *The onset resonant frequencies of noise pole (F_p for the single-pole noise and F_2 for the multiple-pole noise)*:
  1100, 1300, 1500, 1700, 1900, 2100, and 2300Hz.
- *The frequency compression or expansion ratios, which were defined as percent ratios of the frequency compression or expansion against the uncompressed condition*:
  60% and 80%(compressed), 100% (uncompressed), 120% and 140% (expanded).

Procedure The stimuli were presented through binaural head-phones (STAX SR-Λ Signature) in a soundproof room. The presentation level was about 75 dB SPL. The speech

samples were presented in a random order. Subjects were requested to identify the synthetic stimuli as one of the Japanese voiceless stop consonants or voiceless fricative consonants. Each token was presented 20 times for each subject. Three adult subjects participated.

3.2 Results and discussion

Results are shown in **Fig. 6**. The boundary pole frequencies, where the responses of /t/ and /k/ cross in the figure, shifted depending on the frequency-compression or expansion for both noise conditions, the single-pole noise and the multiple-pole noise. The inclination of the shift against the frequency- compression or expansion ratio was steeper for the multiple-pole noise than for the single-pole noise. However, the difference in the inclination seems to be insignificant with regard to the normalization. Rather, it reflects the role of the higher noise formants in the multiple-pole noise condition, which just act as an "weight" on the frequency component constituting a gross spectral shape which indicates a cue for identifying voiceless stop consonants.

4 CONCLUSION

The results of the above experiments suggest that the identification of voiceless fricative consonants between /s/ ans /ʃ/ in frequency compressed speech are mainly based on the absolute noise pole frequency, that is, that the normalization does not occur within voiceless fricative consonants. At the same time, the identification is affected by the vocalic context. The phonetic category boundary between /t/ and /k/, on the other hand, shifted in relation to frequency-compression and expansion independently of the number of prevocalic noise poles. This result suggests that normalization for the compression or expansion of the frequency axis occurs in the identification of voiceless stop consonants.

REFERENCES

[1] S. Sekimoto, S. Kiritani and S. Saito, "Intelligibility of frequency compressed speech in low-pass filtered condition," Ann. Bull. RILP, 14, 181-193, 1980.

[2] S. Sekimoto, "Perceptual normalization of frequency scale," Ann. Bull. RILP, 16, 95-101, 1982.

[3] T. C. Rand, "Vocal tract size normalization in the perception of stop consonants," Haskins Laboratories SR-25/26, 141-146, 1971.

[4] J. May, "Vocal Tract Normalization for /s/ and /ʃ/," Haskins Laboratories SR-48, 67-73, 1976.

[5] O. Kunisaki and H. Fujisaki, "On the influence of context upon perception of voiceless fricative consonants," Ann. Bull. RILP, 11, 85-91, 1977.

[6] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," JASA, 49, 2(Pt. 2), 583-590, 1971.

[7] V. A. Mann and B. H. Repp, "Influence of vocalic context on perception of the [s]-[ʃ] distinction," Perception & Psychophysics, 28 (3), 213-228, 1980.

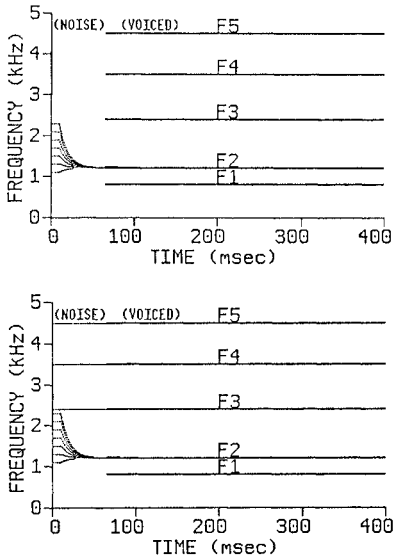

Fig. 5. The time patterns of the control parameters for synthesizing /ta/-/ka/ for the single-pole noise (top) and the multiple-pole noise (bottom).

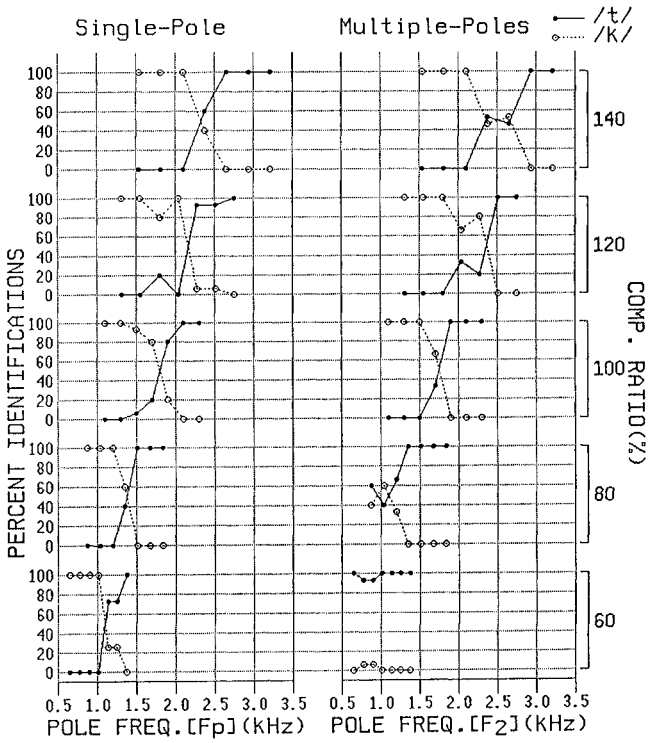

Fig. 6. The identification rates for /t/ and /k/, where the prevocalic noise is approximated by the single-pole noise (left) and the multiple-pole noise (right).