Ann. Bull. RILP, No. 10, 145-156 (1976)

ANALYSIS, RECOGNITION, AND PERCEPTION OF VOICELESS FRICATIVE CONSONANTS IN JAPANESE*

Hiroya Fujisaki and Osamu Kunisaki**

Abstract

A model is described for the spectral characteristics of voiceless fricative consonants of Japanese, based on an equivalent circuit representation of their generation mechanism. The model, together with its three simplified versions, is then evaluated from the point of view of automatic recognition as well as of synthesis of speech. For automatic recognition, spectral models that contain zeros are found to be particularly effective, and their parameters are shown to be sufficient for the complete separation of /s/- and /ʃ/- samples in CV and VCV utterances. On the other hand, perceptual experiments using synthetic stimuli reveal considerably smaller differences between models with spectral zeros and those without zeros.

I. Introduction

While it is true that processings at various linguistic and non-linguistic levels do contribute to resolving certain ambiguities at the acoustic level, and thus to improving the reliability of a system for automatic recognition of connected speech, the amount of improvement is definitely limited, since a certain amount of information lost at the earliest level can never be retrieved from the context. Furthermore, even the most elaborate speech-understanding system would fail to interpolate ambiguities where a human listener will have no difficulty to do so, since it is far beyond the capabilities of present-day computer technology to simulate even a minor part of a whole system of human knowledge to which an ordinary listener has easy access. Thus accurate extraction of the acoustic features of speech and their effective utilization are considered to be most essential as the first step in the automatic recognition of connected speech.

In the case of vowels and vowel-like sounds, it is well established that the formant frequencies, or frequencies of poles of the vocal tract transfer function, are most effective in describing their acoustic characteristics [1]. The importance of formant frequencies has been demonstrated in the synthesis as well as in the automatic recognition of vowels of various languages [2], [3].

The concept of formants has also been extended to the characterization of nasal and lateral consonants [4], [5], where branching of the vocal tract produces not only poles but also zeros, manifested as 'anti-formants' in the spectrum. The situation is similar in most voiceless consonants, for which the major source of excitation is situated midway in the vocal tract and thus produces zeros as well as poles of the transfer function [6] - [8]. Considerably little is known, however, about their actual values as

** Paper presented at the 1976 International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, April 12-14, 1976.

** Department of Electrical Engineering, Faculty of Engineering, University of Tokyo.

well as their effectiveness in the synthesis and the recognition of these sounds, which may be ascribable to the lack of spectral models capable of properly expressing their essential features.

The present paper describes an approach to obtain spectral models that are useful both for synthesis and recogntion of voiceless fricative consonants /s/ and /ʃ/ of Japanese. Based on an equivalent circuit representation of the production mechanism of these sounds, a simplified model is derived for their spectral envelopes. Parameters of the model are then determined from measured spectra by the method of Analysis-by-Synthesis [2]. The model, together with its various simplifications, is evaluated from the point of view of automatic recognition. On the other hand, perceptual validity of these models and their parameters is examined by absolute identification tests and paired comparison tests, with a view to their application to speech synthesis.

II. A Model of Frequency Spectra of Voiceless Fricative Consonants

The excitation source of the voiceless fricative consonants /s/ and /ʃ/ can be considered as the random pressure fluctuation caused by turbulence of the airflow at the outlet of the vocal tract constriction [9], formed by the tongue tip and the upper gum as shown in Fig. 1; the exact conditions for the occurrence and the exact location of such turbulence are not known, however, even for constrictions of much simpler shapes than those found in the vocal tract.

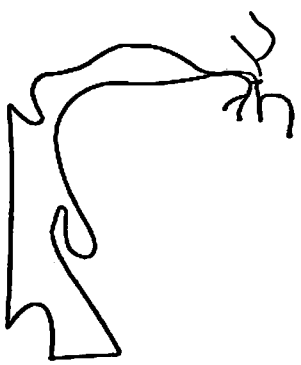

Fig. 1. Midsagittal section illustrating an articulatory configuration appropriate to the fricative consonant /s/ (adapted from Fant [1]).

In order to obtain a clear insight into the general characteristics of sounds produced by such a vocal tract configuration, however, we simplify the situations and regard the entire vocal tract as consisting of three parts, viz., the constriction, the front cavity, and the back cavity, each being approximated by an acoustic tube of uniform cross-sectional area, as shown by the equivalent circuit of Fig. 2.

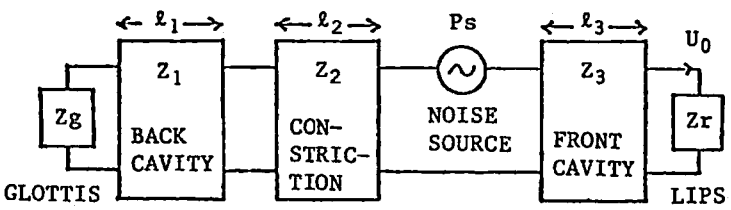

Fig. 2. Equivalent circuit for the production mechanism of voiceless fricative consonants.

By assuming further that the input impedance of the back cavity is negligible as compared to the characteristic impedance of the constriction, and that the radiation impedance from the lips is also negligibly low as compared to the characteristic impedance of the front cavity, and neglecting all losses, the transfer admittance T relating the turbulent noise pressure source to the volume current at the lips can be given by Eq. (1)

$$T = \frac{-j\cos(\omega \ell_2/c)}{Z_2 \sin(\omega \ell_2/c)\cos(\omega \ell_3/c) + Z_3 \sin(\omega \ell_3/c)\cos(\omega \ell_2/c)} , \quad (1)$$

where ℓ_i and Z_i respectively denote the length and the characteristic impedance of the i-th section of the vocal tract, and the suffixes 2 and 3 respectively refer to the constriction and the front cavity.

Figure 3 shows the dependency of pole and zero frequencies of the transfer admittance for the idealized vocal tract configuration on the ratio (ℓ_2/ℓ_3), assuming a constant ℓ_3 and an impedance ratio (Z_2/Z_3) of 50.

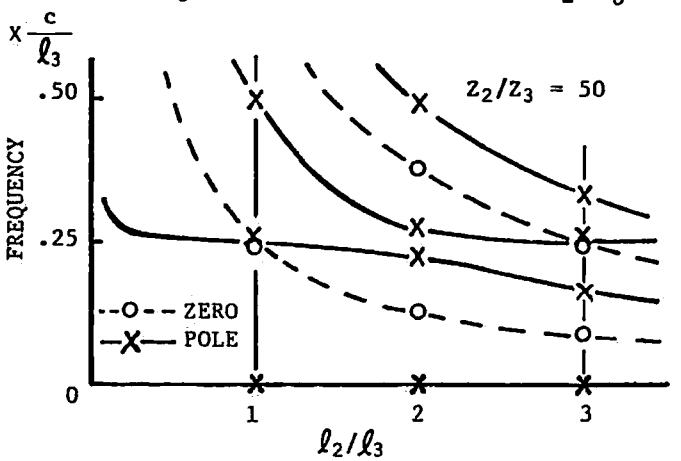

Fig. 3. Pole-zero patterns of the transfer admittance of a simplified vocal tract configuration for production of voiceless fricative consonants.

The transfer admittance is found to possess a pole at the origin, and then one zero and two poles in the ascending order of the frequency, if the ratio (ℓ_2/ℓ_3) lies between 1 and 3, but the order of frequencies of the zero and the next pole is reversed if the constriction is relatively short, and thus the ratio (ℓ_2/ℓ_3) is less than unity.

On the other hand, frequency spectrum of the turbulent noise source, obtained experimentally by inserting a spoiler in a tube of uniform cross-sectional area, indicates a broad peak at a frequency determined by the ratio of the flow velocity and the characteristic dimension of the spoiler[10]. Our preliminary spectral analysis of fricative consonants, however, indicates that it is sufficient to assume a constant slope for the logarithmic power spectrum of the source over the frequency range up to 5 kHz. The radiation transfer impedance relating the airflow at the lips to the sound pressure at one point in the space can be approximately represented by an emphasis characteristic of +6 dB/oct.

These considerations lead to the following model for the actual frequency spectra of /s/ and /ʃ/ [11]:

$$P(s) = K\left[\prod_{i=1}^{m} \frac{(s-s_{zi})(s-s_{zi}^*)}{s_{zi}s_{zi}^*}\right]\left[\frac{1}{s}\prod_{j=1}^{n} \frac{s_{pj}s_{pj}^*}{(s-s_{pj})(s-s_{pj}^*)}\right]s^{1+\alpha}, \quad (2)$$

$$s = 2\pi jf, \quad s_{zi} = 2\pi(jF_{zi} - B_{zi}), \quad s_{pj} = 2\pi(jF_{pj} - B_{pj}),$$

where the first term K represents noise source intensity, the second and the third terms respectively represent contributions of zeros and poles within the frequency range of interest. The last term represents the combined effect of radiation characteristics, higher poles and zeros, and deviation of the noise source spectrum from exactly flat characteristics.

The number of zeros (m) and that of poles (n) except the one at the origin naturally depend on the frequency range of interest, and our preliminary analysis indicates that reasonable approximations to frequency spectra of /s/ and /ʃ/ can be obtained by putting m = 1 and n = 2, suggesting that frequencies of one zero and two poles might be sufficient as parameters for recognition as well as for synthesis of these consonants. From a practical point of view, however, it is also important to know whether or not they are necessary, and, if not, to what extent the model can be simplified. In the following analysis, therefore, the validity of each of the four versions, corresponding to all possible combinations of m = 0, 1 and n = 1, 2 are investigated. For the sake of brevity, the version with m = 1 and n = 2 shall be denoted as Model$_{12}$, as shown in Table 1.

Table 1. Four models for the frequency spectra of voiceless fricatives /s/ and /ʃ/.

Model	Model$_{12}$	Model$_{11}$	Model$_{02}$	Model$_{01}$
No. of zeros	1	1	0	0
No. of poles	2	1	2	1

III. Analysis of Voiceless Fricative Consonants

The validity of the proposed model can be tested by its ability to approximate the actual frequency spectra of voiceless fricative consonants. The speech materials used for this purpose were 60 words of CV and VCV type consisting of all possible combinations of the five Japanese vowels and the two consonants /s/ and /ʃ/, which were uttered by a male speaker. These materials were recorded in an anechoic chamber, low-pass filtered at 9.6 kHz, sampled at 20 kHz with an accuracy of 11 bits per sample, and analyzed by a digital computer.

The first step of the analysis was derivation of frequency spectra of consonantal segments of speech, extracted by a 50-msec hanning window and converted into logarithmic power spectra over the frequency range of 0 to 10 kHz. The frequency range from 0.3 to 5.0 kHz was then divided into 24 frequency bands according to Mel scale and the averaged level within each band was calculated to represent the smoothed spectral envelope.

The second step of the analysis was determination of pole and zero frequencies as well as the overall slope of the spectral envelope of the model for each of these measured spectra by the method of Analysis-by-Synthesis, i.e., by finding their values yielding the best approximation to the measured spectral envelope in terms of the least mean squared error criterion. The Q's of poles and zeros were held constant at their respective values found in preliminary analysis, as listed in Table 2.

Table 2. Values of Q for poles and zeros used in the four versions of the spectral model.

	$Model_{12}$	$Model_{11}$	$Model_{02}$	$Model_{01}$
Q of zero, Qz	2	2	—	—
Q of pole, Qp	4	4	7	12

Examples of such an approximation by each of the four versions of the spectral model are shown in Fig. 4 for one sample of /s/, and in Fig. 5 for one sample of /ʃ/, together with parameter values thereby extracted. The smooth curves in these figures indicate best approximations based on the model, and the symbol + indicates a point on the smoothed envelope of a measured spectrum.

These results indicate that spectral models incorporating a zero ($Model_{12}$ and $Model_{11}$) are much superior to those without a zero ($Model_{02}$ and $Model_{01}$) in the approximation of spectral envelopes of /ʃ/, while their differences are comparatively smaller in the approximation of spectral envelopes of /s/. This fact can be ascribed to the relative proximity of the zero and one of the poles in /s/, which reduces the difference between models with a zero and those without a zero.

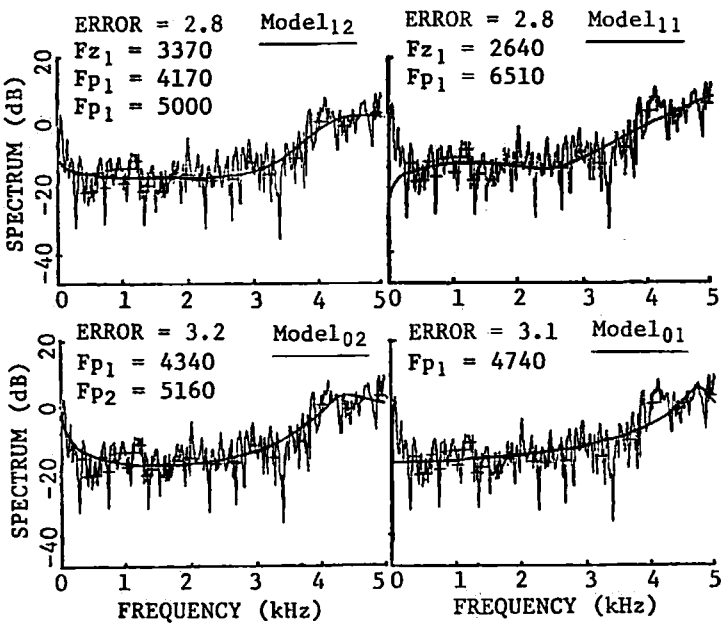

Fig. 4. Comparison of four model versions for a spectral sample of /s/ in /sa/.

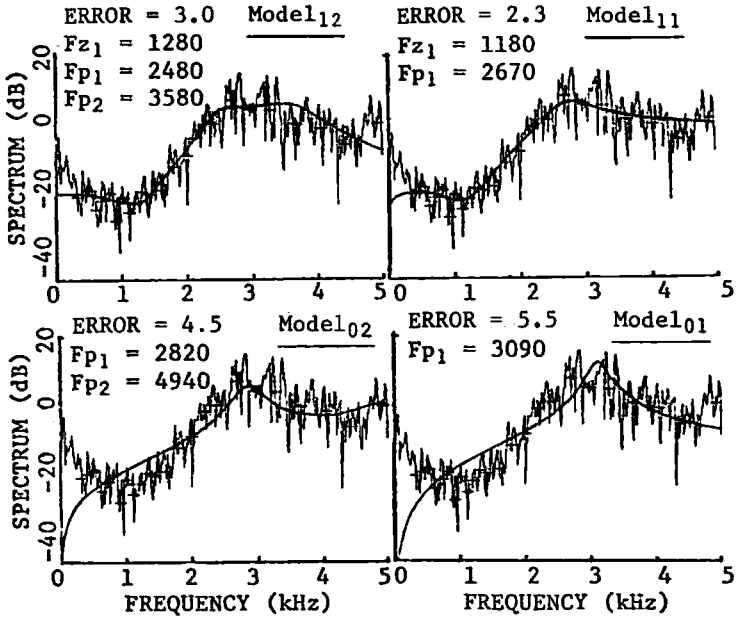

Fig. 5. Comparison of four model versions for a spectral sample of /ʃ/ in /ʃa/.

IV. Classification of /s/ and /ʃ/ in Parameter Space

The close agreement found between some versions of the model and measured spectra of /s/ and /ʃ/ indicates the model's capability of expressing the essential features of these sounds, and suggests the validity of their use in the automatic recognition. In order to evaluate various versions of the model and their parameters from this point of view, the optimum linear discriminant function was determined for each version of the model by the following procedure.

First, a vector in the parameter space is determined such that the ratio of inter-class variation to intra-class variation is maximized when sample points belonging to the two classes /s/ and /ʃ/ are mapped onto the vector. The vector can be obtained as the eigen vector $\mathbf{a}$ corresponding to the largest eigen value of the following matrix:

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{M_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_{i.})(\mathbf{x}_{ij}-\bar{\mathbf{x}}_{i.})^t\right\}^{-1}\left\{\sum_{i=1}^{n}M_i(\bar{\mathbf{x}}_{i.}-\bar{\bar{\mathbf{x}}}_{..})(\bar{\mathbf{x}}_{i.}-\bar{\bar{\mathbf{x}}}_{..})^t\right\}$$

where $\mathbf{x}_{ij}$ denotes the vector corresponding to j-th sample of i-th class, $\bar{\mathbf{x}}_{i.}$ denotes the mean vector of i-th class, $\bar{\bar{\mathbf{x}}}_{..}$ denotes the overall mean vector of all the classes, M_i denotes the number of samples belonging to i-th class, and i = 1 and 2 respectively corresponding to /s/ and /ʃ/. Secondly, by assuming normal distributions of the sample points of each class on vector $\mathbf{a}$, a decision threshold b is determined to minimize the probability of classification error. The optimum linear discriminant function in the

parameter space is then given by the hyperplane

$$F(x) = ax - b \quad . \tag{3}$$

As an example, distributions of /s/- and /ʃ/-samples on the F_{z1}-F_{p1} plane, extracted by means of $Model_{12}$, are shown in Fig. 6, together with the linear discriminant function determined by the above-mentioned procedure. In this example, complete separation of /s/- and /ʃ/-classes, each consisting of 30 samples, is found to be possible even without the third parameter, F_{p2}.

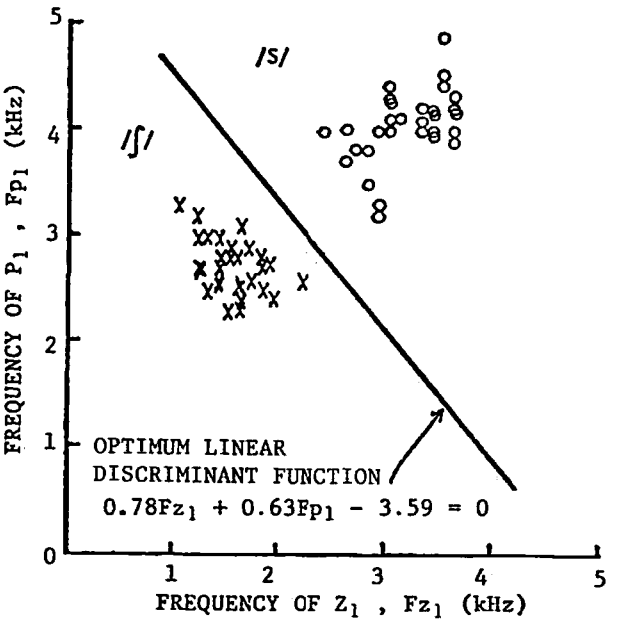

Fig. 6. Distribution of /s/- and /ʃ/-samples in F_{z1}-F_{p1} plane of $Model_{12}$ and the optimum linear discriminant function.

The same procedure can be applied to determine the relative importance of each of the parameters. The result shows that F_{z1} is the most important among the parameters, and is followed by F_{p1}, in separating /s/ and /ʃ/.

Analysis of variabilities in the extracted values of F_{p1} and F_{z1} also indicates that they are caused mainly by contextual influences of adjacent vowels. In particular, the greater variability of F_{p1} is found to be mostly due to the dependency of the front cavity length (ℓ_3) on the vowel immediately following the fricative, ℓ_3 being larger and thus causing a decrease in F_{p1} when the vowel is /o/ or /u/ as compared to the case when the vowel is /a/ or /e/. On the other hand, the smaller variability of F_{z1} reflects the relative stability of the constriction length (ℓ_2) against such coarticulatory influences. Contextual dependency of F_{p1} and F_{z1} values is schematically illustrated by Fig. 7.

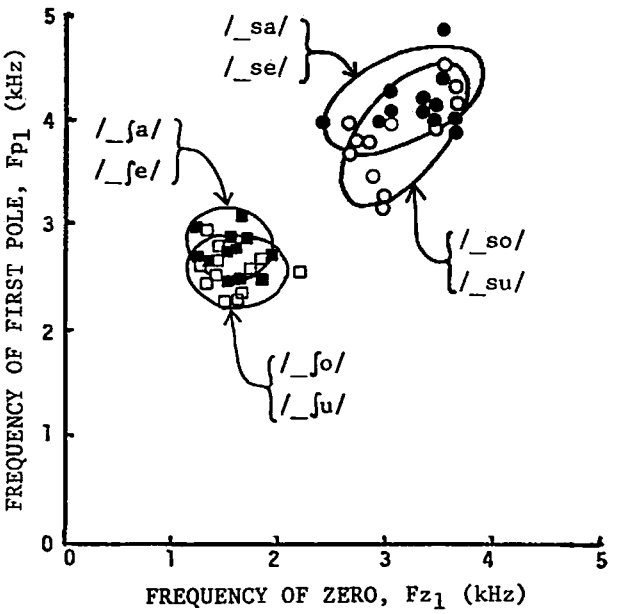

Fig. 7. Influence of following vowels on extracted parameters of fricative consonants /s/ and /ʃ/ in the word-medial position.

Models with reduced number of poles and zeros are also evaluated in the same way, and the results, expressed in terms of the percentage of correctly classified samples, are summarized in Table 3. The table also lists the root mean squared errors in dB per point on the spectral envelope of /s/ and /ʃ/, for each of the four versions of the spectral model. These results indicate that 100% correct recognition is possible by Model_{12} and Model_{11}, while errors occur in the case of parameters obtained by models without a zero.

Table 3. Comparison of four model versions for fricative spectra in terms of average matching error (dB/point) in Analysis-by-Synthesis and the percentage of correct classification by optimum linear discriminant function.

Model		Model_{12}	Model_{11}	Model_{02}	Model_{01}
Average errors (dB/point)	/s/	2.61	2.72	2.79	3.10
	/ʃ/	3.05	2.99	4.41	5.15
Recognition rate (%)		100 (Fz_1-Fp_1-Fp_2)	100 (Fz_1-Fp_1)	97 (Fp_1-Fp_2)	92 (Fp_1)

V. Synthesis and Perception of /s/ and /ʃ/ [12]

Though the foregoing analysis indicates the importance of the frequency of a spectral zero in the automatic recognition of /s/ and /ʃ/, its perceptual significance has yet to be examined. It should also be investigated whether the linear discriminant function, determined to minimize the probability of classification error in the parameter space of each of the four versions of the model, corresponds to the boundary of perceptual categorization by human listeners. These questions can best be answered by identification tests of synthetic stimuli. From the point of view of speech synthesis, it is also important to know to what extent the model can be simplified without seriously affecting the perceptual quality of speech sounds synthesized by the model. For this purpose, paired comparison tests can be used to establish a psychological scale of naturalness for the four versions of the model.

Identification Tests

The stimuli used for the identification tests were synthetic CV syllables generated by computer simulation of a terminal-analog speech synthesizer. The vowel synthesizer consisted of a buzz source followed by five cascaded digital pole circuits and an emphasis circuit representing the radiation characterisitic, while the consonant synthesizer consisted of a random noise generator followed by one zero and two pole circuits, each of which could be bypassed if necessary.

In the present study, the vowel was restricted to /a/ with linear transitions of 75 msec in the first two formant frequencies, while its higher formant frequencies were held constant, as shown in Fig. 8. These formant patterns were simplifications of those measured in utterances of a male speaker. The fundamental frequency of the vowel was held constant at 150 Hz, and the intensity of the stationary part of the consonant relative to that of the stationary part of the vowel was fixed at -12 dB, based on preliminary listening tests. The Q's of poles and zeros of fricative consonants were made identical to those used in the analysis.

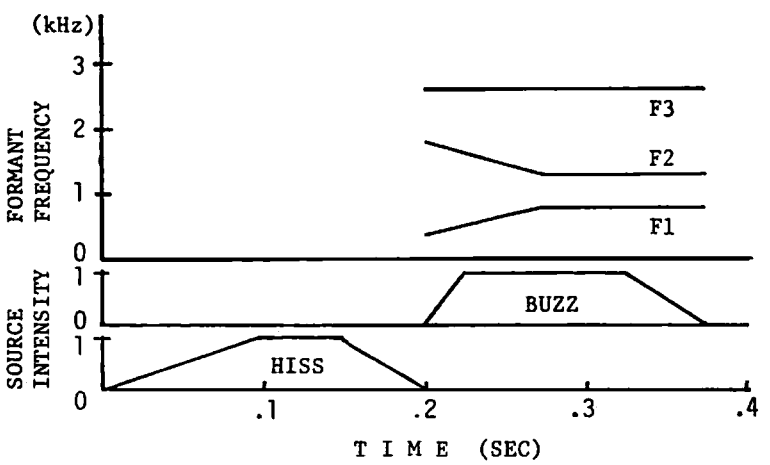

Fig. 8. Time chart of vowel formant frequency and source intensity control for the synthesis of consonant-vowel syllables.

The consonantal parts of synthetic stimuli used for the identification tests were selected at points in the parameter space so as to represent both typical sounds as well as intermediate sounds for /s/ and /ʃ/. Figures 9(a) to (d) respectively indicate the stimuli for each of the four versions of the model. These stimuli were presented in random order at intervals of 4 sec for forced judgment. The subjects were two male adults with normal hearing; at least 20 judgments on each of the stimuli were made. The solid curves in Fig. 9(a) to (c) indicate averaged results for the two subjects in the form of equi-probability contours of /s/-judgment, while Fig. 9(d) shows the results for Model$_{01}$ as an identification curve. The optimum linear discriminant function is also shown in each figure by a broken line. The magnitude and the direction of the gradient vector in each figure respectively indicate absolute and relative importance of individual parameters. While rather close agreement is found in Model$_{12}$ and Model$_{01}$ between the optimum linear discriminant function and the perceptual boundary (50%-judgment contour), there is marked discrepancy between the two in other models. For example, the first pole frequency (F_{p1}) is found to be perceptually more important than the zero frequency (F_{z1}) in Model$_{11}$, while F_{z1} is found to be more important than F_{p1} for automatic recognition.

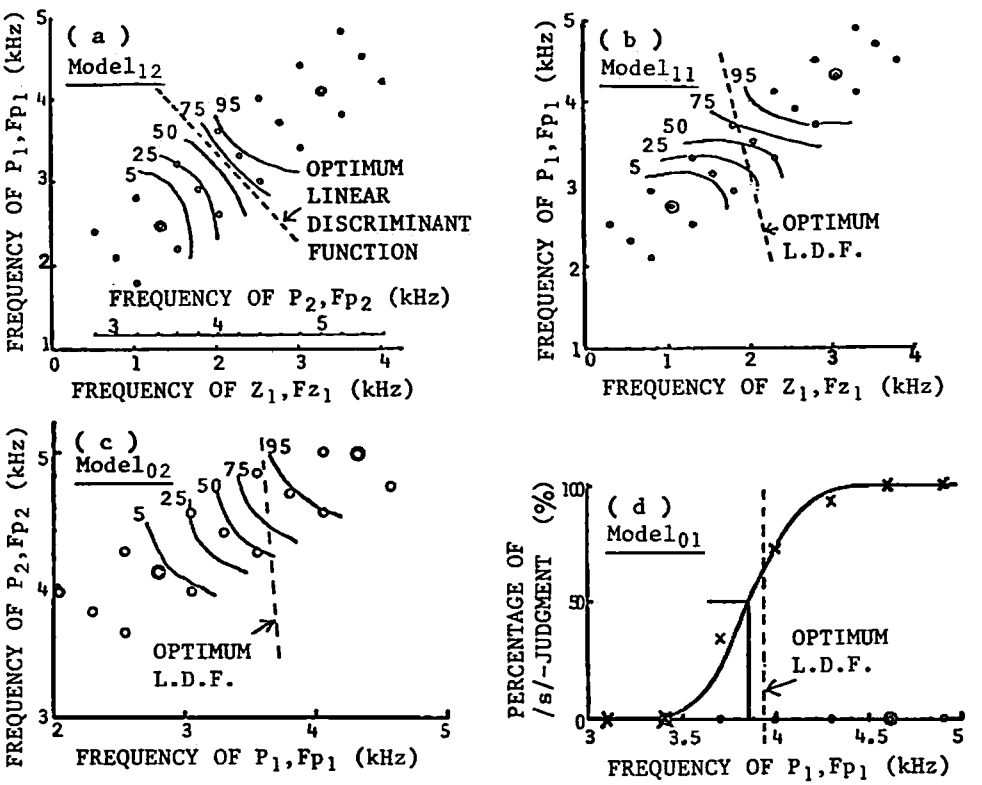

Fig. 9. Stimuli and results of identification tests of fricative consonants synthesized by the four model versions.

Paired Comparison Tests

The stimuli used for the paired comparison tests were generated by using a natural utterance of the word /asa/ or /aʃa/ as the basis, replacing only the fricative portion by its closest synthetic approximation based on each one of the four model versions. The original natural utterance was also used to serve as the reference, thus resulting in a total of five different tokens for /asa/ and also five for /aʃa/. Separate tests were conducted for /asa/ and /aʃa/.

All possible pairs of the five tokens, including identical pairs, were prepared with an inter-stimulus interval of 1 sec and presented in random order at intervals of 4 sec. The subjects were eight male adults with normal hearing, who were forced to answer which of the paired stimuli they judged to be more natural than the other. The results of the eight subjects were pooled and processed by Thurstone's method [13] to derive a scale of relative naturalness of the synthetic and natural tokens of /asa/ and another scale of /aʃa/. In each case, the natural utterance served as the reference point.

Figure 10 shows the comparison of naturalness of the four model versions for the /s/-sound in /asa/ as well as for the /ʃ/-sound in /aʃa/. The ordinate in each case is the naturalness scale thus derived with the natural utterance as the origin, and a negative number on this scale indicates a loss of naturalness expressed in terms of the mean standard deviation of all judgments. These results indicate that Model_{12} and Model_{01} show practically no difference from the natural utterance, while some degradation is perceivable if we use other model versions. It is particularly useful to know that, with a proper selection of parameter values, the simplest, single-pole model can be preceptually almost as acceptable as the natural fricative sounds and their closest approximation, i.e., Model_{12}.

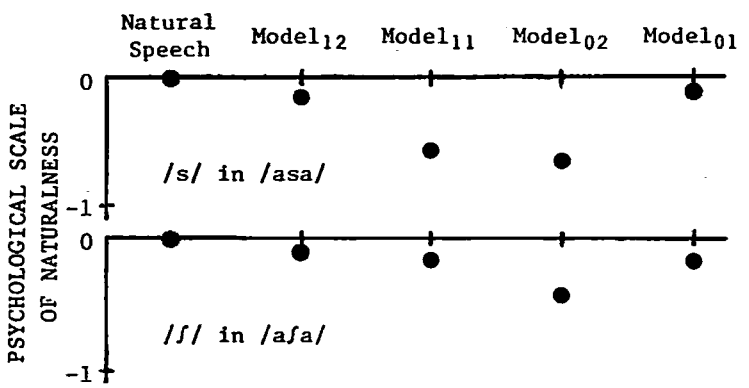

Fig. 10. Comparison of naturalness of four versions of synthetic fricatives with a natural utterance as reference.

VI. Conclusions

A model has been presented for the spectral characteristics of voiceless fricative consonants /s/ and /ʃ/, and its parameters have been derived from measured spectra of these sounds. The model, together with three simplified versions, has been evaluated from the point of view of automatic recognition as well as synthesis of speech. While the frequency of a spectral zero is found to be quite important in separating /s/ and /ʃ/, and hence validates the use of such a model for automatic recognition, its perceptual significance is considerably smaller than expected. Further study is in progress both on inter-speaker differences of the spectral parameters and their normalization for automatic recognition, as well as perceptual study of parameter normalization for /s/ and /ʃ/.

References

1. Fant, G. (1960), Acoustic Theory of Speech Production, Mouton.
2. Fujisaki, H., N. Nakamura, and K. Yoshimune (1970), "Analysis, Normalization and Recognition of Sustained Japanese Vowels," J. Acoust. Soc. Japan, 26, 152-154.
3. Fujisaki, H. and B. Kim (1973), "Analysis and Recognition of Korean Vowels," Annual Rept., Eng. Res. Inst. Fac. Eng., University of Tokyo, 32, 227-232.
4. Fujimura, O. (1962), "Analysis of Nasal Consonants," J. Acoust. Soc. Amer., 34, 1865-1875.
5. Kato, Y. (1965), "Analysis of Lateral Consonants," Rec. Spring Meeting, Acoust. Soc. Japan, 73-74.
6. Heinz, J. M. (1958), "Model Studies of the Production of Fricative Consonants," MIT R. L. E. Quart. Progr. Rept., No. 50, 146-149.
7. Heinz, J. M. and K. N. Stevens (1961), "On the Properties of Voiceless Fricative Consonants," J. Acoust. Soc. Amer., 33, 589-596.
8. Heinz, J. M. (1961), "Analysis of Fricative Consonants," MIT R. L. E. Quart. Progr. Rept., No. 60, 181-184.
9. Flanagan, J. L. and L. Cherry (1969), "Excitation of Vocal-Tract Synthesizers," J. Acoust. Soc. Amer., 45, 764-769.
10. Stevens, K. N. (1971), "Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations," J. Acoust. Soc. Amer., 50, 1180-1192.
11. Fujisaki, H. and O. Kunisaki (1975), "Analysis and Recognition of Voiceless Fricative Consonants in Japanese," J. Acoust. Soc. Japan, 31, 741-742.
12. Kunisaki, O., T. Matsuo and H. Fujisaki (1976), "Perceptual Study of Voiceless Fricative Consonants Using Synthetic Stimuli," Rec. Spring Meeting, Acoust. Soc. Japan, 327-328.
13. Thurstone, L. L. (1961), The Measurement of Values, Chicago University Press.