Ann. Bull. RILP (1983) No. 17, 83-96

NORMALIZATION OF THE SPEAKER DIFFERENCE IN VOWEL PERCEPTION

Sotaro Sekimoto

1. Introduction

It is well known that the phonemic quality of vowels is characterized by formant frequencies, especially by the lowest two formants. However, large variation is found in vowel formants between classes of speakers, that is, men, women and children. It is sometimes found, for example, that the position of a certain vowel on the F_1-F_2 plane uttered by a male speaker overlaps with that of a different vowel uttered by a female speaker.[1-3] Although such remarkable variation is presents in formant frequencies, the vowel that the talker intends is correctly recognized by the listener. Several scaling algorithms as well as factors have been proposed to normalize formant frequency variations.[4,5] but a satisfactory solution has not yet been obtained. As for perceptual studies, there are scarcely any.[6] In the perceptual process, it is supposed that some normalization mechanism exists to compensate for these formant variations. It has not been clarified, however, what property or cues in speech plays a role in such a normalization mechanism.

The purpose of the present study is to clarify the normalization mechanism. In a previous paper, by the present author, the presence of a framework effect and an effect for fundamental frequency were discussed.[7] In the present study, three experiments were conducted.

2. General Method

Experimental design

In the experiments, the following were examined.

(1) The effect of the interaction between the fundamental frequency and the spectrum envelope.
(2) The hypothesis that the ratio between the formant frequencies is the essential determinant of the phonemic vowel quality.
(3) The role of the frequency components above F_2.

Identification tests were carried out on the vowel stimuli in which the stimulus conditions were systematically altered. Frequency compressed or expanded vowels, processed by the method to be described below, were used as stimuli, and the range of the frequency compression or expansion ratio, in which the processed vowels were correctly recognized as the originals, were compared for the respective experimental conditions. Here, the assumption was made that the condition under which the widest range of recognition was obtained could be regarded as that in which normalization works most effectively.

Description of stimuli

Frequency compressed or expanded speech is speech in which the frequency axis is linearly compressed or expanded. The relation between the frequency spectrum envelope and the frequency compression or expansion ratio is schematized in Fig. 1. The relative position of vowels in the F_1-F_2 plane, shown in Fig. 2, can be similarly altered by changing the frequency compression or expansion ratio. In this figure, different vowels with different frequency compression or expansion ratios overlap, as similarly observed for male and female speech. In the previous study by the present author, however, it was observed that frequency compressed or expanded vowels were correctly identified as their originals over a wide range of frequency compression or expansion ratios.[7] This fact suggests that the same normalizing mechanism, which works on natural male and female speech, works on frequency compressed or expanded speech as well. In the present experiment, therefore, frequency compressed or expanded speech was used as stimuli.

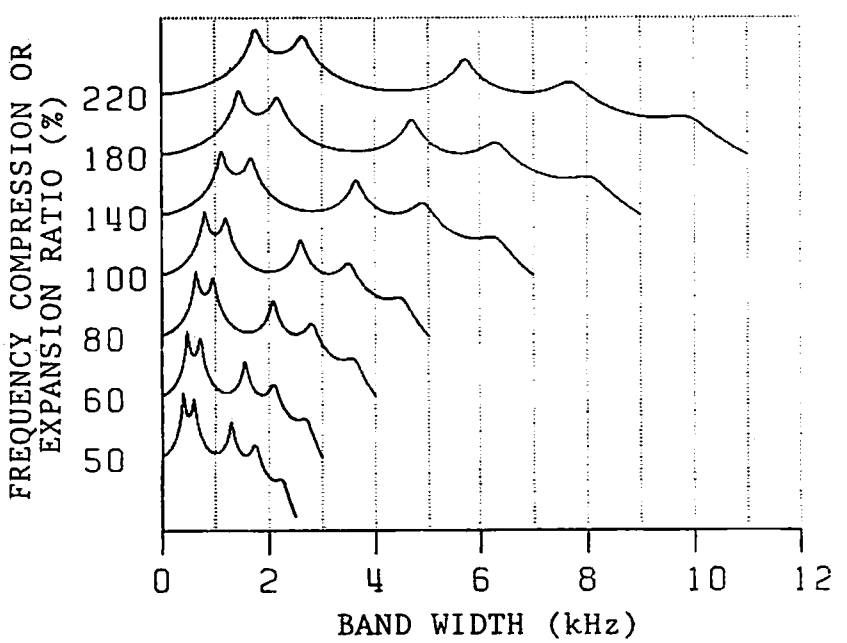

Fig. 1 *Frequency spectrum envelopes for speech with various frequency compression or expansion ratios.*

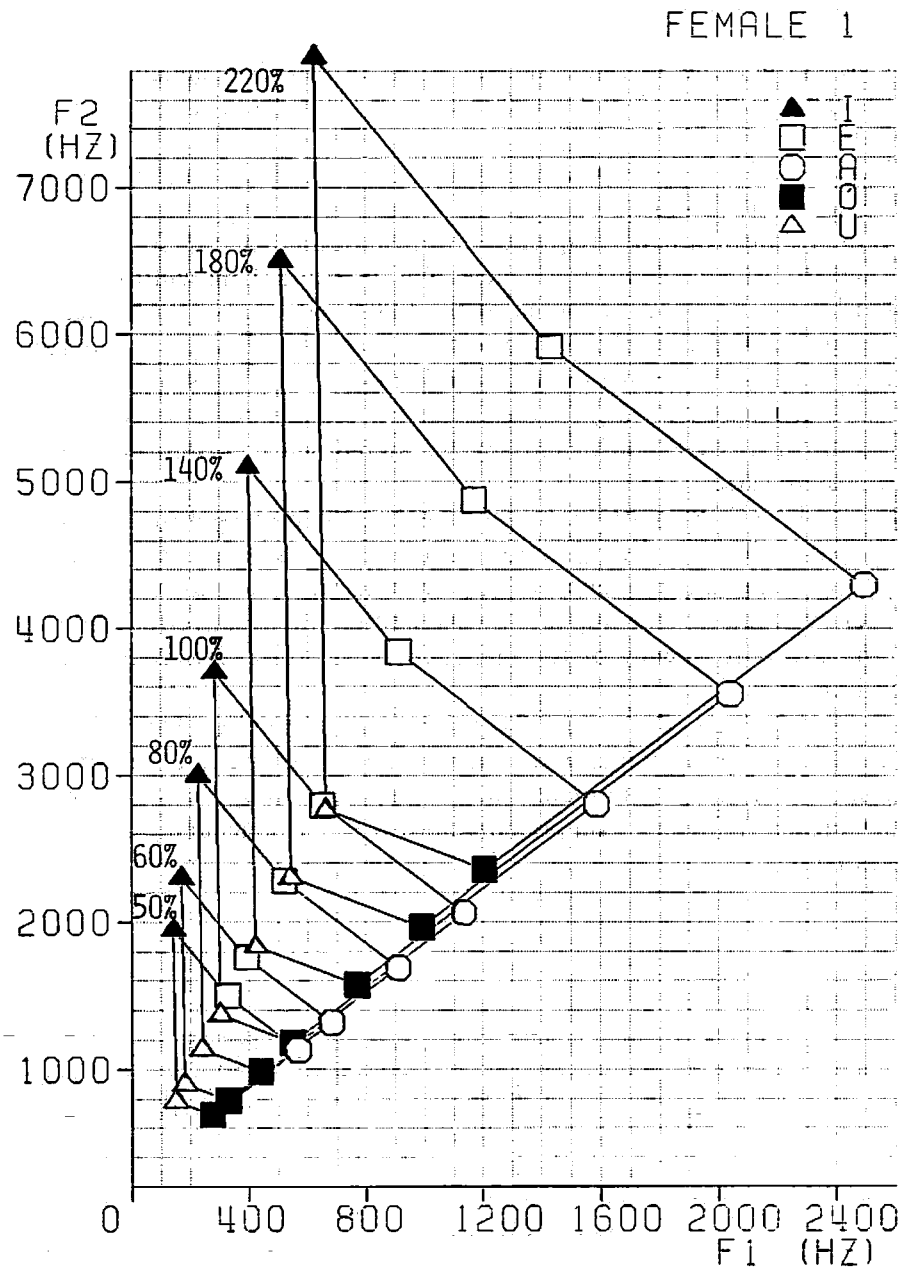

Fig. 2 *F_1-F_2 representation of the five Japanese vowels uttered by one female speaker. The pentagonal outlines encompass the same frequency expansion or compression ratio.*

Processing method for the stimuli

The frequency compression or expansion was carried out on a PDP-11/34 computer with software. The flow-diagram of the speech processing is shown in Fig. 3. The speech processing was performed by a PARCOR speech analysis-synthesis technique. The original speech signal was sampled at 10 kHz and digitized with a 10-bit accuracy. The digitized speech signal was subjected to a PARCOR analysis to extract the LPC coefficients (a_0-a_{12}) and the fundamental frequency (F_0), through a Hamming window 30 msec in length. The analysis was repeated at 5 msec intervals.

In Exps. I and III, these parameters were sent directly to the PARCOR synthesizer (SW was turned to a).

In Exp. II, the formant frequencies were modified in the following way (SW was turned to b): Formant frequencies (F_1-F_6) and bandwidths (B_1-B_6) were calculated by solving the polynomial of the LPC coefficients. The formant frequencies were modified at the specified ratio to be described later, whereas bandwidths were kept unchanged. From the modified formant frequencies and bandwidths, the new LPC coefficients (a'_0-a'_{12}) were reconstructed. These LPC coefficients were converted into PARCOR coefficients by the step-down recursion process, and then sent to the synthesizer.

With the PARCOR synthesizer, the frequency compressed or expanded speech was synthesized. The frequency compression or expansion was simply achieved by changing the sampling frequency of the synthesizer. The output signal from the synthesizer was fed to the digital low-pass filter; then converted into an analog signal. The cutoff frequency of the digital filter was set at different values in Exps. I/III and Exp. II.

In Exps. I and III, the cutoff frequency was set at the "Through" condition, that is, at the same condition as when the signal did not feed to the filter.

In Exp. II, the cutoff frequency was varied, depending on the F_2 value, to remove the higher frequency components above F_2. The accuracy of the D/A converter was 10 bits.

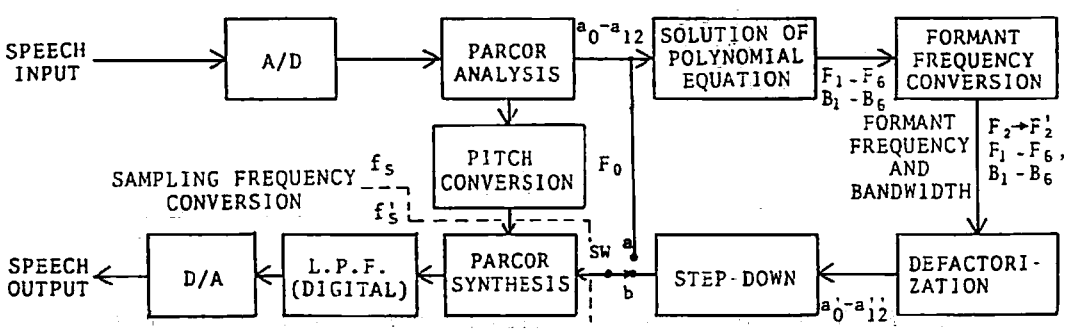

Fig. 3 *Flow diagram of the speech processing based on the PARCOR speech analysis-synthesis technique.*

3. Experiment I

3.1 Method

Stimuli

Frequency compressed or expanded speech in which the condition of the fundamental frequency was varied was used as stimuli. The original speech samples were five steady state Japanese vowels. A total of four male and four female speakers were used for the different experiments. The speech materials were processed by the method described above and used as stimuli.

Stimuli conditions

The following stimuli conditions were examined as variables.

(1) Frequency compression or expansion ratios:
60, 80, 100, 140, 180, 220 and 260%.
(2) Fundamental frequency (F_0)
a) Original F_0 : F_0 was set at the same value as the original.
b) Proportional F_0 : F_0 was raised or lowered in proportion to the frequency compression or expansion ratio, respectively.
c) Middle F_0 : F_0 was set at the middle value between a) and b).

Subjects

One male subject participated. He had some experience in experiments of this kind.

Procedure

The experiment was carried out under on-line computer control. The speech stimuli were presented to the subject through a loudspeaker in front of the subject directly from the computer. The subject was tested in a quiet room. He sat in front of the computer display terminal and the tablet digitizer. The subject was told that he had to identify the stimulus as one of a five Japanese vlwels. His responses were stored in the computer after he pointed to the choices on the tablet with a stylus pen.

Each condition of the fundamental frequency was examined in separate sessions. A session was composed of a set of 100 stimuli, which contained 20 repetitions of each vowel. The stimuli were presented successively in random order. The time period between tokens was 4 sec, and if a response was not returned to the computer, the presentation of the next token was delayed until the response was returned. The presentation level of the stimuli was set at a comfortable level, about 75 dB SPL.

3.2 Results and remarks

The results, for one female speaker, are shown in Fig. 4. The abscissa shows the frequency compression or expansion ratio. Complete identification was found in the range from 80 to 180%, independent of the F_0 condition. For the ratios outside this range, the intelligibility was affected by the F_0 condition. The intelligibility

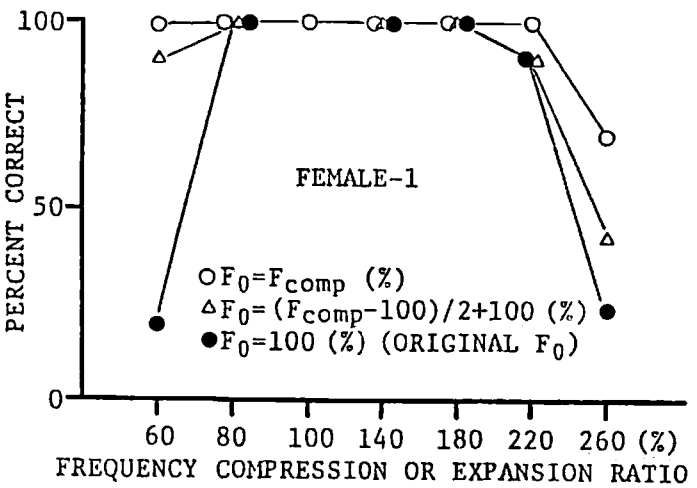

Fig. 4 *Vowel intelligibilities for frequency compressed or expanded speech under various fundamental frequency conditions.*

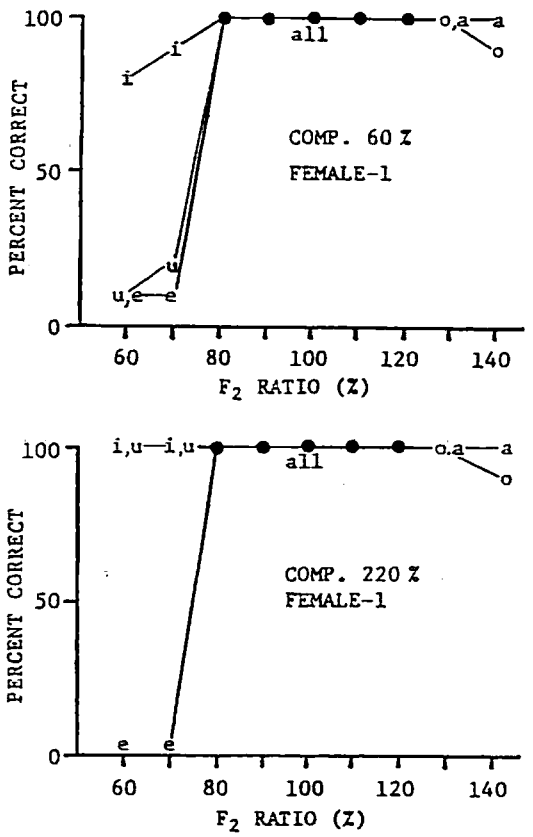

Fig. 5 *Vowel intelligibilities for frequency compressed or expanded speech with various F_2 ratios. The upper figure shows the results with a frequency compression ratio of 60%. The lower figure shows the results with a frequency expansion ratio of 220%.*

was the highest for the Proportional F_0, lower for the Middle F_0, and lowest for the Original F_0. A similar tendency was observed for the stimuli by the other speakers as well.

It was observed that vowels were correctly recognized over a wide range of frequency compression or expansion ratios. In frequency compressed or expanded vowels, the ratios of the formant frequencies were kept constant independent of the frequency compression or expansion ratio. The results of this experiment suggest that the condition under which a certain value is maintained is important for the normalization of the formant differences between speaker classes. Moreover, a certain relation between the fundamental frequency and the spectrum envelope plays an auxiliary role in the normalization. In the next experiment, the role of the ratio of the formant frequencies was explored in further detail.

4. Experiment II

4.1 Method

Stimuli

To exmine the condition under which a constant formant ratio was maintained, intelligibility scores were compared for vowels, in which the ratios between F_1-F_2 were systematically varied. This alteration in the formant frequency ratio could be easily achieved by shifting the formant frequency alone. In the present study, only F_2 was varied.

Stimulus conditions

Frequency compressed or expanded speech in which the second formant frequency was varied was used as stimuli. The original speech samples were five steady state Japanese Vowels. One male and two female speakers with a normal speaking capacity were used. The following conditions were examined as variables.

(1) F_2 ratio
The second formant frequency was raised or lowered from that of the original. The value of the F_2 was varied within the range of
60 – 120% for /i/, /u/ and /e/
80 – 140% for /a/ and /o/
in 10% steps.
The range was determined for each vowel so that the F_2 did not overlap with adjacent formants.

(2) The frequency compression or expansion ratios were
60, 80, 100, 180, 200 and 220%.

(3) The fundamental frequency was changed in proportion to the frequency compression or expansion ratio.

Subjects

Three paid subjects participated. They were not the same as the speakers.

Stimulus tapes

Frequency compressed or expanded vowels were synthesized from the original speech on the PDP-11/34 computer, and recorded on a PCM tape. The conditions for the F_2 ratio and the frequency compression or expansion ratio were examined in separate sessions. A session was composed of a set of 100 stimuli, which had 20 repetitions of each vowel. The stimuli were recorded in random order. The time period between tokens was 2 sec. A pure tone marker was inserted every 10 tokens.

Procedure

The subjects were tested in a sound proof room. They were instructed to identify each stimulus as one of the five Japanese vowels. They responded to the on-line computer by pointing to choices on a digitizer tablet with a stylus pen. The responses were automatically totaled on the computer, and the intelligibility scores and confusion matrixes were calculated. The stimuli were presented to one ear over DR-305 headphones at a comfortable intensity, about 74 dB SPL.

4.2 Results and remarks

Fig. 5 shows the relationship between the vowel intelligibility and the F_2 ratio. The upper figure shows the results for the frequency compression or expansion ratio of 60%. The vowels /a/ and /o/ were almost always correctly recognized for the entire F_2 range of 80–140%. As for the other vowels, /i/, /u/ and /e/ were correctly recognized in the F_2 range from 80 to 120%. In the F_2 range below 80%, the intelligibility for /i/ was high, whereas the intelligibility for /u/ and /e/ were significantly low.

The lower figure shows the results for the frequency compression or expansion ratio of 220%. All of the vowels except /e/ were almost always correctly recognized for all ranges of the F_2 ratio. As for the vowel /e/, it was correctly recognized within the F_2 range of 80% to 120%. In the F_2 range of 60 to 70%, however, the intelligibility scores for /e/ were zero. A similar tendency was found for all other frequency compression or expansion ratios above 100% as well.

It was observed that the perception of vowels is considerably stable even if the second frequency is varied. This result suggests that frequency normalization cannot be completely explained by the formant frequency ratio alone. It was noted that the identification of each vowel did not change in all the ranges of F_2 unless the frequency was expanded or compressed. This implies that a relative vowel space is maintained even if the frequency is compressed or expanded.

5. Experiment III

5.1 Method

In Experiment III the effect of the frequency components above F_2 on frequency normalization was examined. The range of the frequency compression or expansion ratio in which the vowels were correctly identified was compared for two conditions: when the higher frequency components were inserted and when they were deleted.

Stimuli

The frequency compressed or expanded speech was synthesized from original speech samples. Five steady state Japanese vowels were used as the speech samples. Two male and two female speakers were used. The higher frequency components above F_2 were deleted by a digital low-pass filter as mentioned in section 2.1. The filter was inserted before D/A conversion to maintain a similar relation between the frequency characteristics of the low-pass filter and the spectrum envelope of the speech, independent of the variation in the frequency compression or expansion ratios. Fig. 6 illustrates their relationship. The cutoff frequency of the low-pass filter was set at a value 100 Hz lower than the lowest point of the valley of the spectrum envelope between the second and the third formants. The gain of the low-pass filter was decreased by 30 to 40 dB within 300 Hz above the cutoff frequency. The values of 100 and 300 Hz mentioned above were varied in proportion to the frequency compression or expansion ratio. The conditions of the frequency compression or expansion ratio were 50, 60, 80, 100, 160, 180, 200, 220, 240 and 260%. The fundamental frequency was raised or lowered in proportion to the frequency compression or expansion ratios.

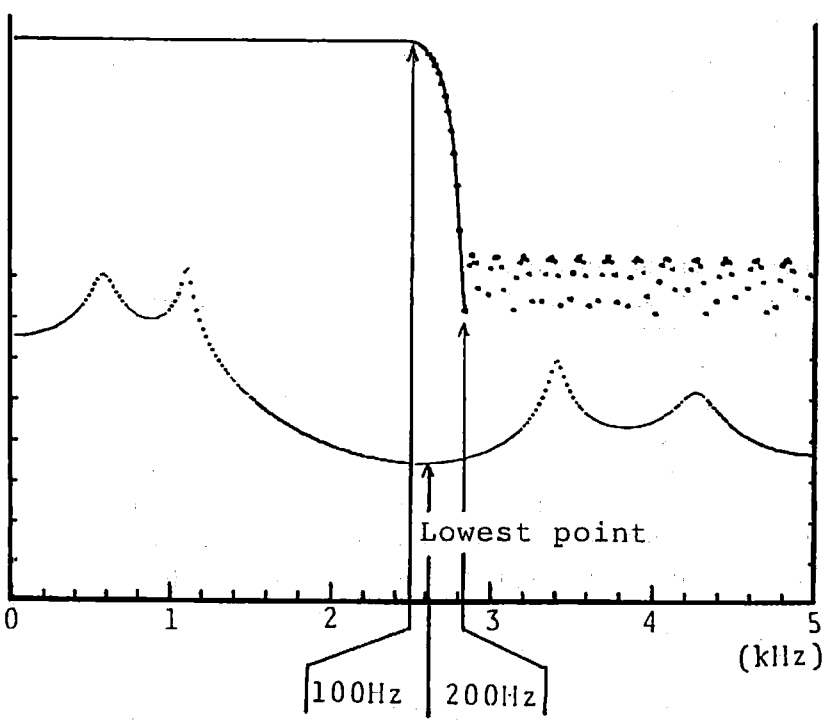

Fig. 6 *Frequency relationship between the filter characteristic and the formant envelope.*

Subjects

Two paid adult subjects participated. They were not the same as those used in Exp. I or II.

Procedure

Tests were carried out for each speaker and each vowel, separately, in independent sessions. In one session, the speaker and the vowel were fixed and the frequency compression or expansion ratio was varied. Each ratio was repeated 10 times in a session. This method was adopted to avoid the framework effect of the frequency compression or expansion condition on the vowel identification.[7] The general procedure was the same as for Experiment II.

5.2 Results and remarks

Fig. 7 shows the results for the vowels /o/ and /u/ uttered by a female speaker. In these figures, the results for the two conditions of the higher frequency components are compared.

It was observed for both /o/ and /u/ that the range of the frequency compression or expansion ratio in which the vowel was correctly identified was wider if the higher frequency components were inserted than if they were deleted. The difference in the range between the two conditions was wider for /u/ than for /o/. This difference is considered to have been caused by the difference in the relative intensity of the higher frequency components compared to the lower ones between the vowels /o/ and /u/.

Fig. 8 shows the results for the same vowels uttered by a male speaker. In this figure, results similar to those in Fig. 7 can be observed.

It can be assumed that the higher frequency components play a role in the normalization of formant differencies between speaker classes. However, it is necessary to confirm the assumption that phonemic quality is not changed even if the higher frequency components are removed. Accordingly, a supplementary experiment was conducted.

The intelligibility of vowels in which the F_2 ratio was varied were compared in the conditions where the higher frequency components were removed and not removed. The method was similar to that used in Exp. II. Fig. 9 shows the results for a female speaker, in which the vowel intelligibility is plotted against the F_2 ratio. A similar tendency was found in the intelligibility curve for all of the vowels, irrespective of the presence/nonpresence of the higher frequency components. As for the vowels /a/ and /u/, the intelligibility was always 100%; therefore, these results are omitted. These results are not similar to those in Exp. II due to speaker differences.

It is now confirmed that the higher frequency components above F_2 play a role in the normalization of the formant differences between speaker classes.

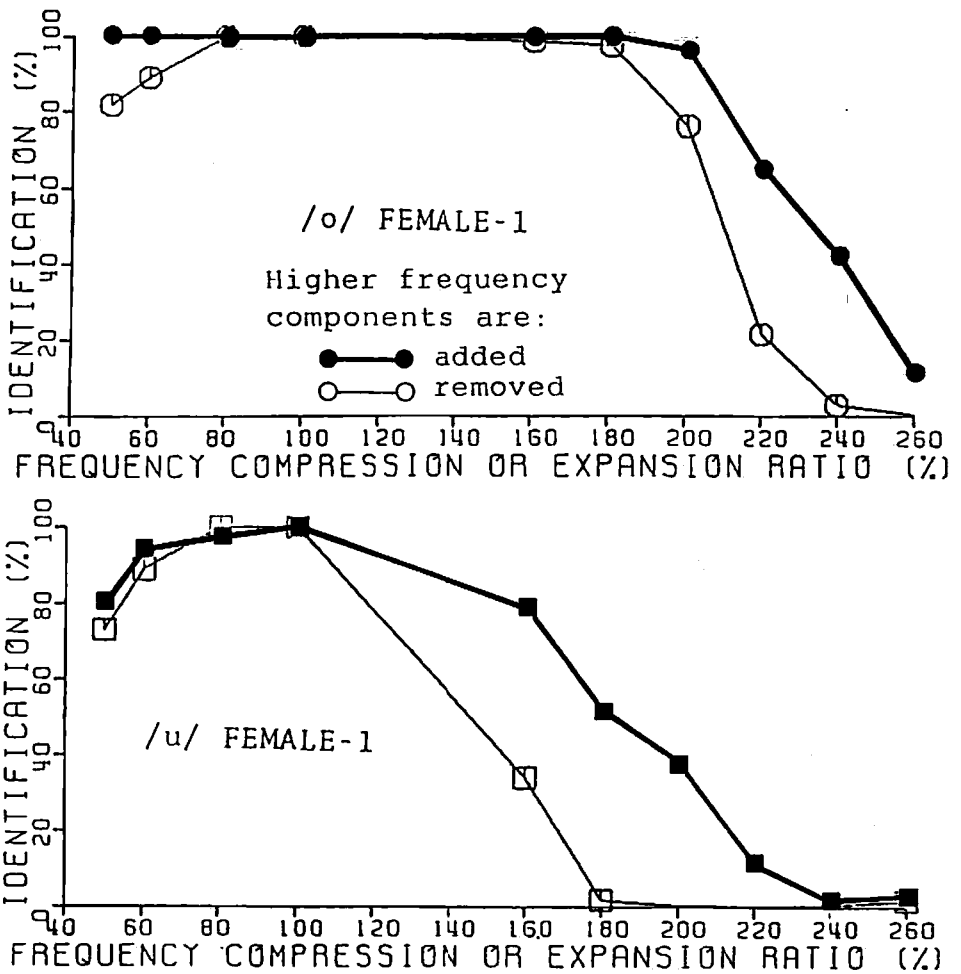

Fig. 7 *Vowel intelligibilities for frequency compressed or expanded speech with the higher frequency components added or removed. The results for a female voice are shown.*

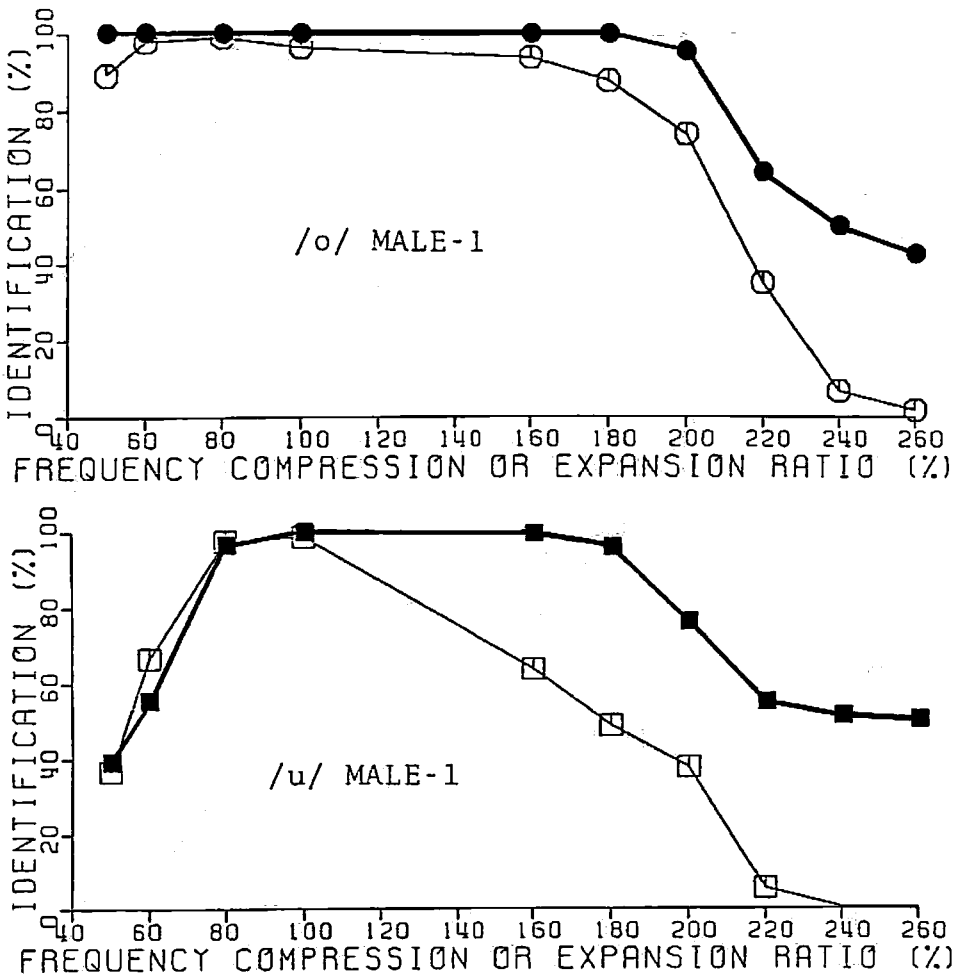

Fig. 8 *Vowel intelligibilities for frequency compressed or expanded speech with the higher frequency components added or removed. The results for a male voice are shown.*

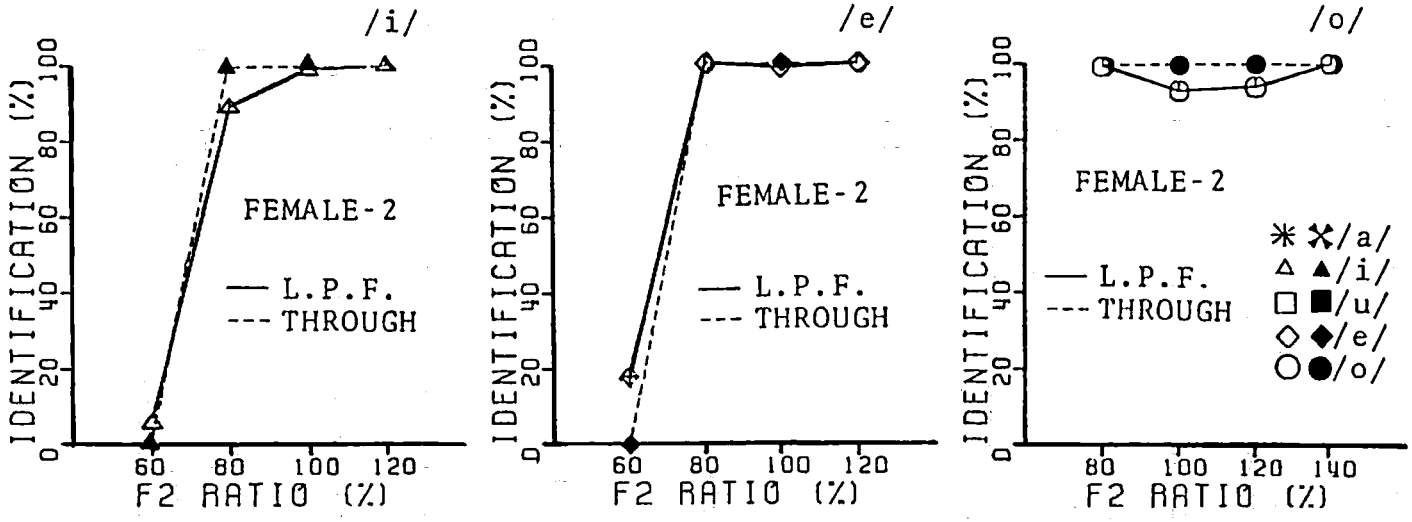

Fig. 9 *Identification of vowels with various F_2 ratios. The results are compared for conditions where the higher frequency components are added and those where they are removed.*

6. General Remarks

In this study, frequency compressed or expanded speech was used as stimuli to simulate the differencies in formant frequencies between speakers. The use of the frequency compressed or expanded speech made possible a systematic approach to speaker difference in the perceptual experiment. This new method for the perceptual study of the normalization of speaker differences in vowel formant space should be explored further.

The outcome of the present study can be summarized as follows.

(1) The ratios between the formant frequencies were essential to the perceptual normalization of speaker differences in vowel formant space, but a relative judgment within the vowel space was still possible.
(2) A certain relation between the fundamental frequency and the spectrum envelope was important.
(3) The higher frequency components played an auxiliary role.

As for the roles of pitch and the higher formants in vowel perception in an ordinary formant space, i.e.. which is not compressed nor expanded, a study has already been reported by Fujisaki and Kawashima.[6] Results (2) and (3) above agreed with theirs. In addition, these roles were clearly maintained even when the frequency scale was changed.

As other factors might exist, further research on this topic should be pursued.

Acknowledgment

This study was supported in part by a Grant-in-Aid for Scientific Research (No. 57780024) from the Japanese Ministry of Education, Science and Culture.

References

1. Peterson, G.E. and H. Barney (1952): Control Methods Used in a Study of the Vowels. J. Acoust. Soc. Am., 24, 175-184.
2. Fant, G. (1959): Acoustic Analysis and Synthesis of Speech with Application to Swedish, Ericsson Technics, No. 1.
3. Kasuya, H., H. Suzuki and K. Kido (1968): Changes in Pitch and first Three Formant Frequencies of Five Japanese Vowels with Age and Sex of Speakers, J. Acoust. Soc. Japan 24, 355-364.
4. Fant, G. (1966); A note on vocal tract size factors and non-uniform F-pattern scalings, STL-QPSR 4/1966, 22-30.
5. Fujisaki, H., N. Nakamura and K. Yoshimune (1970); Analysis, normalization and recognition of sustained Japanese vowels, J. Acoust. Soc. Japan 26, 152-154.
6. Fujisaki, H. and T. Kawashima (1968): The Roles of Pitch and Higher Formants in the Perception of Vowels, IEEE Trans. Audio Electroacoust., AU-16 (1), 73-77.
7. Sekimoto, S. (1982): Perceptual normalization of frequency scale, Ann. Bull. RILP, 16. 95-101.