Ann.Bull.RILP(1985)No.19, 65-74

THE ROLE OF CONTEXT IN FREQUENCY NORMALIZATION IN VOWEL IDENTIFICATION

Sotaro Sekimoto

1. Introduction

It is well known that the acoustic structure of speech varies from one speaker to another, especially across males, females and children. It is not resolved how the listener compensates for these variations and recognizes the correct message that the speaker intends. The present author reported on experimental studies to explore the perceptual cues to calibrate the frequency variation in vowel speech using synthesized speech material in which the frequency axis was linearly expanded (Sekimoto 1982; Sekimoto 1983). It was found that the identification of a frequency expanded vowel was affected by other vowels concurrently presented in the same test session and a so-called framework effect was found. It was also found that the relation between the fundamental frequency and formant frequencies was significant. In these studies, the frequency expanded, vowel stimuli were presented one by one in isolation. It is known that a similar phenomenon in the identification of vowels occurs when vowels are embedded in a carrier sentence uttered by different speakers (Ladefoged and Broadbent, 1957). It is believed that this phenomenon implies the existence of a normalization process in the temporal domain. It is not clear, however, what information in the frame sentence influences the recognition of the embedded vowels.

In the present study, as candidates for an effect on identification, the following were investigated.
(1) The number of moras before the test vowel.
(2) Dependency on the vowel class.
(3) The formant frequencies (related to the size of the vocal tract) and/or the formant frequencies of the carrier sentence.

2. Method

Outline of experiments

The five Japanese test vowels (/i/, /e/, /a/, /o/, and /u/), which had the voice quality of a male or a child, were embedded in carrier sentences which had the voice quality of a male, a female, or a child, and the identification scores for the test vowels were measured and compared. The following nine sentences were used as frame sentences.
(1) /kore o _ to yuu/ ("We call it _").
(2) /kore wa _ to yuu/ ("We call it _").
(3) /ii _ da/ ("It's good _").
(4) /aa _ da/ ("Ah!, it is _").
(5) /ee _ da/ ("Yes, it is really _").
(6) /oo _ da/ ("Oh!, it is _").

(7) /a _ da/ ("Ah!, it is _").
(8) /e _ desu ka/ ("Eh ?, is it _ ?").
(9) /o _ da/ ("Well, it is _").

Sentences (1) and (2) had three moras in the initial part of the context and had a contrast in the phoneme just before the test vowel. Sentences (3)-(6) had two moras in the initial part, and a different type of vowel. Sentences (7)-(9) had a mora with the same type of vowel in the initial part as (4)-(6), respectively. /I _ da/ was not used as it is meaningless as a sentence in Japanese.

Experimental conditions

In natural speech uttered by different speakers, significant individual variations are found between speakers, which seems to be irrelevant to the normalization of the differences in an overall formant structure resulting from variation in vocal tract size, such as utterance timing, and local variation in formant frequencies. The effects of these individual variations have to be avoided. In the present study, in order to avoid these effects, speech of female and child voice quality was synthesized from the speech uttered by a male talker, using the PARCOR (LPC) speech analysis-synthesis technique by altering the sampling frequency at the synthesis stage from that of the analysis stage. The ratio of the sampling frequency at the synthesis stage to that at the analysis stage will be referred to as the "frequency expansion ratio" hereafter. The frequency expansion ratios which gave the voice quality of male, female and child were 1.0, 1.4, 2.0, respectively, with the fundamental frequency was raised in the same proportion.

The conditions for the experiments are shown in Table 1. The frequency expansion ratios and the fundamental frequency ratios used for the context sentences and the test vowels in Exps. 1, 2 and 3 are shown. In Exps. 1 and 2, the voice quality of the test vowel was fixed as male and child ones, respectively, and the effect of the carrier sentences with various voice qualities (male, female and child) was observed. In Exp. 3, the effect of the fundamental frequency of the carrier sentence was examined. The voice quality of the test vowel was fixed as child's as in Exp. 2, and only the fundamental frequency was varied while maintaining the formant frequencies as those of a child ones.

Materials

Carrier sentences with each vowel were recorded on a PCM tape recorder in an anechoic room. An adult male of Tokyo dialect served as a speaker. The recorded speech was subjected to PARCOR speech processing. The flow-diagram of the speech processing is shown in Fig. 1. The speech signal was subjected to a PARCOR analysis for the PARCOR coefficients and the fundamental frequency. These parameters were sent to the synthesizer to generate the speech whose frequency components were linearly expand-

Table 1. Experimental conditions

		FREQUENCY EXPANSION RATIO	FUNDAMENTAL FREQUENCY RATIO
Exp. 1	I	1.0	1.0
	CONTEXT II	1.4	1.4
	III	2.0	2.0
	TEST VOWELS	1.0	1.0
Exp. 2	I	1.0	1.0
	CONTEXT II	1.4	1.4
	III	2.0	2.0
	TEST VOWELS	2.0	2.0
Exp. 3	I	2.0	1.0
	CONTEXT II	2.0	1.4
	III	2.0	2.0
	TEST VOWELS	2.0	2.0

ed upward. The frequency expansion was made by altering the sampling frequency at the synthesizer; thereforr, the sampling frequency at the output of the synthesizer was different for different frequency expansion ratios. It was difficult, however, to connect the synthesized speech wave with that for different sampling frequencies. Thus, at the next stage, the sampling frequency conversion was made to make the sampling frequency at the D/A output uniform.

The synthesized version of the sentence was, then divided into three segments: the initial part of the carrier sentence (referred to as "head" hereafter), the embedded test vowel, and the trailing part of the carrier sentence (referred to as "tail" hereafter) as shown in Fig. 2. These segments with various frequency expansion ratios were combined to make a test sentence, in which the test vowel was embedded within a carrier sentence with various frequency expansion ratios. The head segment and the tail segment extracted from the same synthesized sentence were used as a pair. The test vowels extracted from sentence (1) were used for test sentences (1), (2), and (3), and those from sentence (7) were used for test sentences (4)-(9), in order to reduce the interaction which would exist in an original synthesized sentence between the test vowel and the carrier sentence (except for sentence (1)). The duration of the pauses between the head segment and the test vowel and between the test vowel and the tail segment were fixed equal to that of /a/ in order to prevent use as a cue in identifying the vowels.

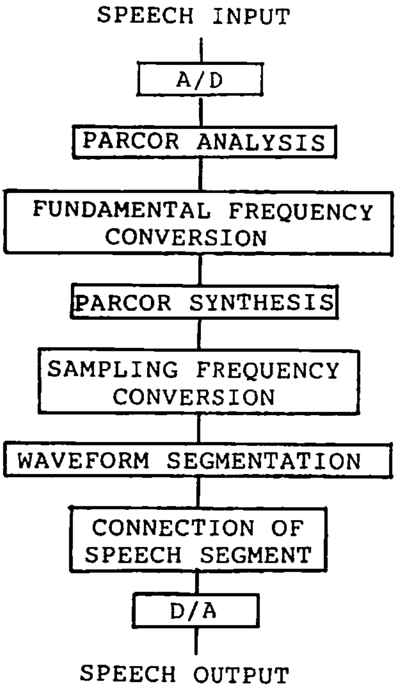

SPEECH INPUT
A/D
PARCOR ANALYSIS
FUNDAMENTAL FREQUENCY CONVERSION
PARCOR SYNTHESIS
SAMPLING FREQUENCY CONVERSION
WAVEFORM SEGMENTATION
CONNECTION OF SPEECH SEGMENT
D/A
SPEECH OUTPUT

Fig. 1 Flow diagram for the speech processing.

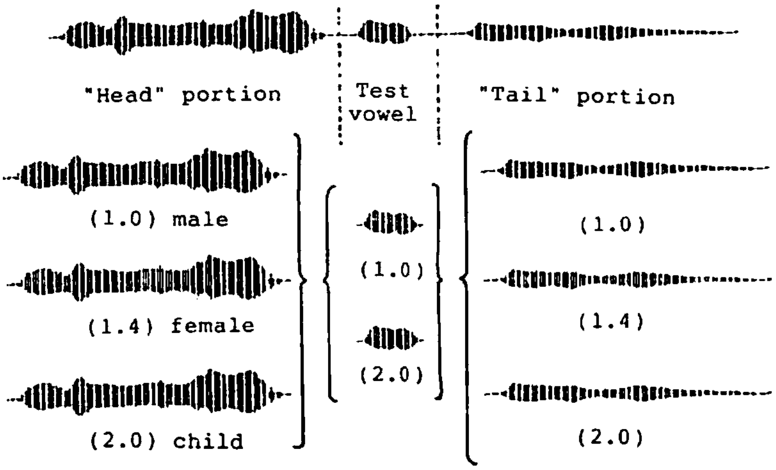


Fig. 2 Segmentation of the synthesized sentence into three portions. Connections of these portions with various frequency expansion ratios were used as simulus tokens. Head and tail portions with the same frequency expansion ratio were used as pairs. Values in parentheses show the frequency expansion ratios.

The A/D conversion of the original speech was made at a sampling frequency of 10 kHz with 12 bit accuracy. The order of the PARCOR analysis was 12. The analysis was repeated at 5 msec intervals with a Hamming window 30 msec in length. The sampling frequency at the D/A output was 20 kHz.

Procedure

Each experiment was carried out in separate sessions. The test sentences with various context conditions were presented in random order. The time period between sentences was 4 sec. A session was composed of a set of 75 sentences, which contained 5 repetitions of each sentence. Each session was repeated 10 times so that the number of presentations for each sentence became 50 in total. The speech stimuli were presented to the subject through a loudspeaker in a sound proof room. The presentation level was about 75 dB SPL. An adult male subject participated.

3. Results and remarks

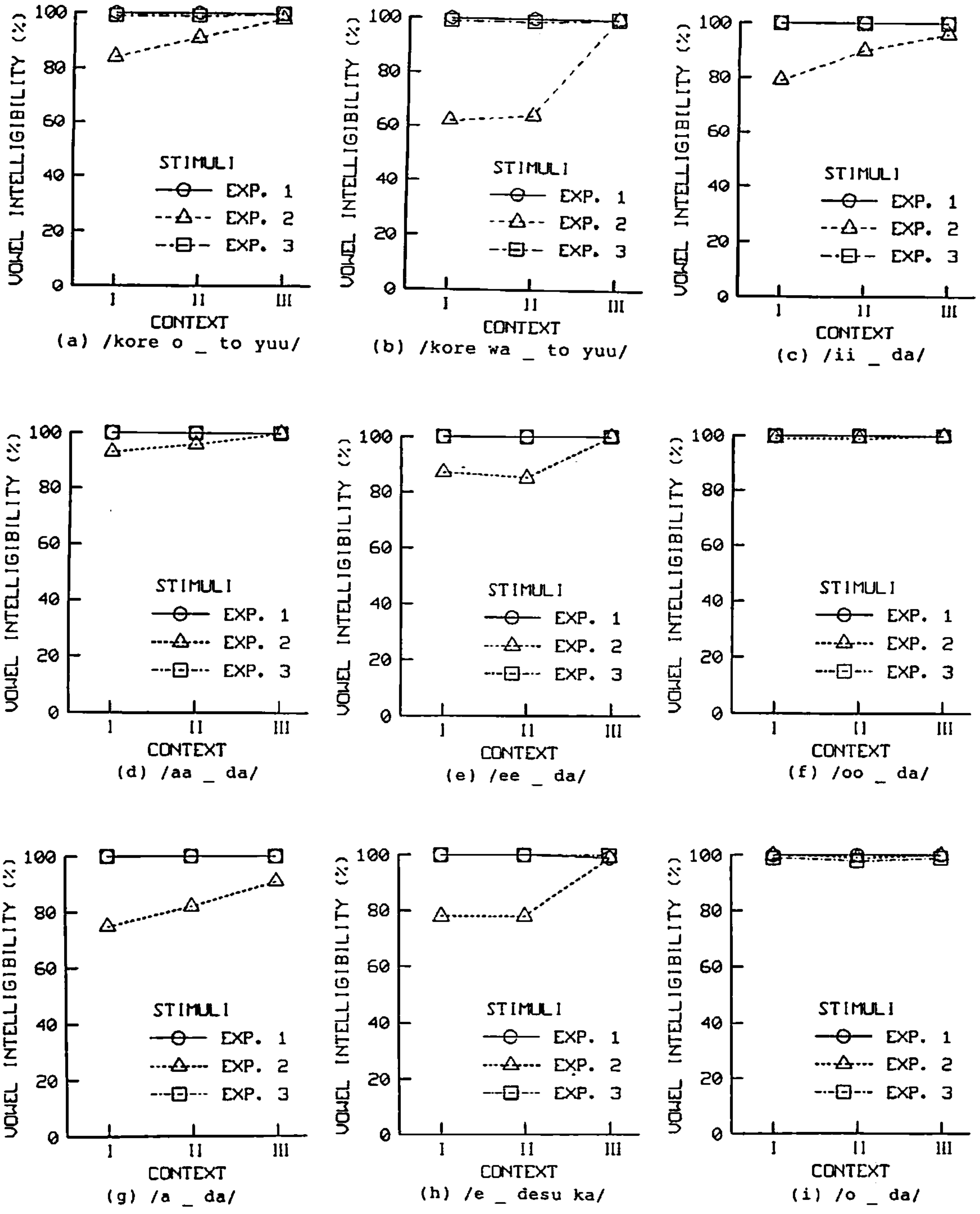
The vowel intelligibility for each test sentence is shown in Fig. 3 (a)-(i). The ordinate shows the averaged percent identification of the five vowels. The abscissa shows the context conditions shown in Table 1.

The intelligibility scores for Exp. 1 were 100%, regardless of the context conditions, for all test sentences.

The intelligibility scores for Exp. 2 deteriorated for the context conditions I and II, but not for the context condition III, for every sentence except sentences (6) and (9). That is, the identification of the test vowels was affected by the context if the voice qualities of the test vowels and the context were different. It was noted that this effect was found even if the number of the preceding vowel was one as in sentences (7) and (8).

The intelligibility scores for Exp. 3, however, were 100% or nearly 100%, for every context condition and sentence. Since the difference in experimental conditions between Exp. 2 and Exp. 3 existed only in the fundamental frequency, this result means that the identification of the test vowels was affected by the context of the formant structure alone, not by the fundamental frequency.

Comparing the results for Exp. 2 for the sentences (1) and (2) (Fig. 3(a) and Fig. 3(b)), the intelligibility score for (1) was greater than that for (2), or the extent of the effect was different for the phoneme just before the test vowel. Comparing the results for Exp. 2 among the sentences (3), (4), (5), and (6), and among (7), (8), and (9), it can be seen that the extent of the effect of the preceding vowel was different among the vowel classes. Comparing the results for Exp. 2 with the number of moras in the head segment (or the length of the vowels preced-



(a) /kore o _ to yuu/

(b) /kore wa _ to yuu/

(c) /ii _ da/

(d) /aa _ da/

(e) /ee _ da/

(f) /oo _ da/

(g) /a _ da/

(h) /e _ desu ka/

(i) /o _ da/

Fig. 3 Vowel intelligibilities for the embedded test vowels of respective test sentences. Figures (a)-(i) show the results for the sentences (1)-(9), respectively.

ing the test vowels) between sentences (4) and (7) and between (5) and (8), the vowel intelligibilities were greater for the head segment with one mora.

The deteriorations in vowel intelligibility in Fig. 3 were mainly caused by a confusion in the identification of the test vowels /o/ and /u/. The ratios of identification for the test vowels /o/ and /u/ are shown in Figs. 4 and 5, respectively. In these figures, the results for sentences (1), (2), and (3) are shown. The test vowels /o/ and /u/ were mainly confused with /a/ and /e/, respectively. Similar results were obtained for other sentences.

The direction of the confusion does not seem to contradict that expected from the relative formant position of vowels which have different frequency expansion ratios. The vowel position on the F_1-F_2 plane extracted from an isolated steady-state vowel uttered by the same speaker is shown in Fig. 6. Vowels with the same frequency expansion ratio are connected by a line. Supposing that the frequency expansion ratio of the carrier sentence is 1.0 and that of the test vowel is 2.0, for example, the smallest polygon will be assumed as a framework when the initial part of the context is presented, and when the test vowel is presented in succession, the vowels /o/ and /u/ will be identified in the relative position in the framework as /a/ and /e/, respectively.

4. Conclusion

In order to elucidate the role of the framework in the temporal domain for the perceptual normalization of speaker differences, what information in a preceding speech segment affects the identification of the vowels has been investigated using synthesized vowels with the voice quality of a male or a child embedded in sentence contexts which had the quality of a male, a female or a child. The following results were obtained.
(1) The identification of the embedded test vowels was not affected regardless of the context condition if the voice quality of the test vowels was male.
(2) If the voice quality of the test vowels was that of a child one, the identification score for the test vowels, and especially for the vowels /o/ and /u/, deteriorated when the voice quality of the context did not coincide with that of the test vowel. On the other hand, the identification score did not deteriorate when the voice quality of the context was equal to that of the test vowel. It was concluded that the preceding speech segments functioned as a framework for the frequency normalization toward the successive vowels in a time domain.
(3) The formant structure of the context, not the fundamental frequency, functioned as a normalization cue if the normalization effect was found in (2).
(4) The effect of the context on the identification of the test vowels was found even when the preceding speech segment was only one syllable.
(5) The extent of the deterioration in (2) varied among vowel

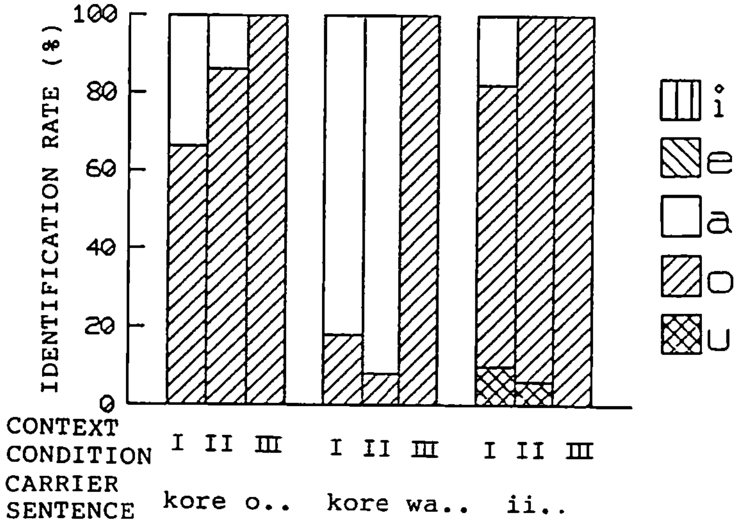


Fig. 4 Identification ratio for the test vowel /o/ in the test sentences (1), (2), and (3).

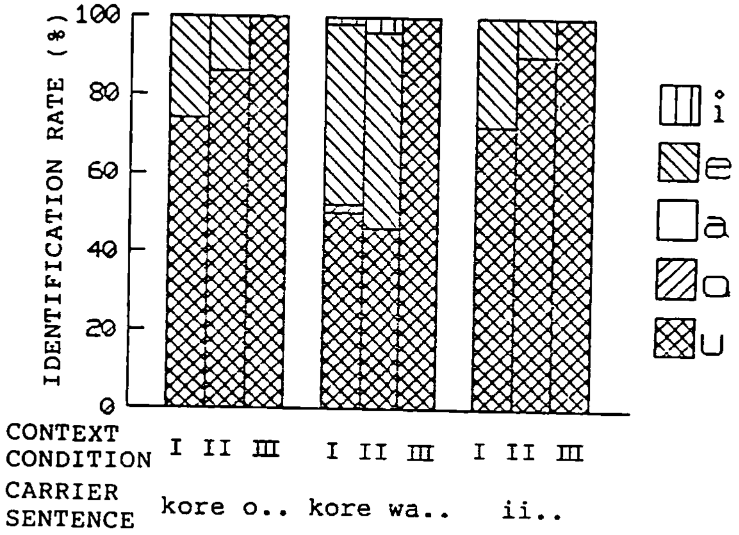


Fig. 5 Identification ratio for the test vowel /u/ in the test sentences (1), (2), and (3).

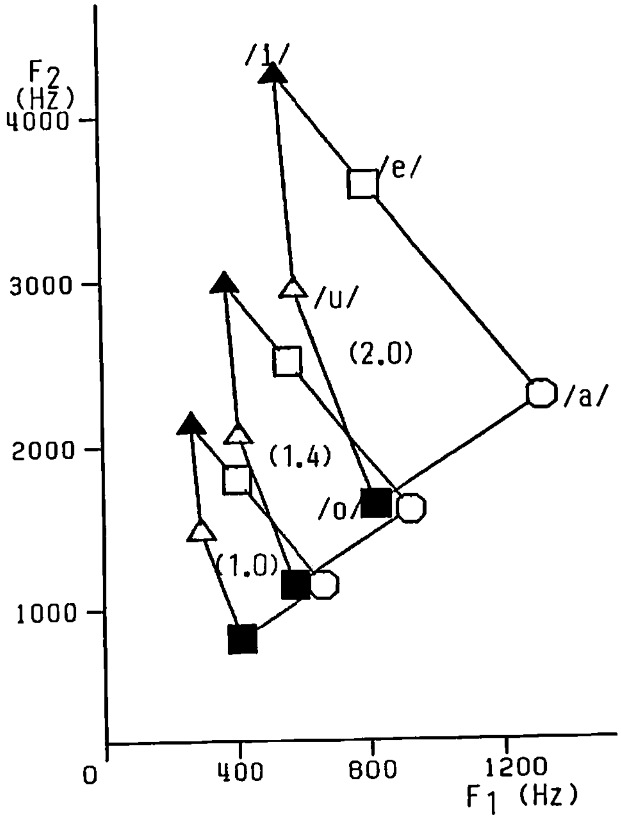


Fig. 6 F_1-F_2 representation of the five Japanese vowels uttered by a male speaker. Vowels with the same frequency expansion ratio are connected with a line. Values in parenteses indicate the frequency expansion ratios.

classes. It was not resolved, however, whether this dependency on vowel class was caused by the so-called context effect (Fry et al., 1962) within the framework and/or an effect inherent in the normalization process.

It is supposed that a common mechanism of the perceptual process might exist between the context effect and the normalization effect. A study on their relationship is now in progress.

References

Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman (1962); The identification and discrimination of synthetic vowels. Language and Speech 5, 171-189. Op.ti 0 Ladefoged P. and D. E. Broadbent (1957); Information Conveyed by Vowels. J. Acoust. Soc. Am., 29, 98-104.

Sekimoto S. (1982); Perceptual normalization of frequency scale. Ann. Bull. RILP, 16, 95-101.

Sekimoto S. (1983); Normalization of the speaker difference in vowel perception. Ann. Bull. RILP, 17, 83-96.