TR-I-0053

On the unit selection measure for speech synthesis by rule using multiple synthesis units

複合合成単位を用いる規則音声合成における単位選択尺度について

Katsuo Abe and Yoshinori Sagisaka
安部勝雄、匂坂芳典

November, 1988

Abstract: In speech synthesis by rule, we have already proposed a synthesis scheme using non-uniform speech synthesis units to obtain the optimum synthesis unit sequence for a desired output speech. In this paper, vowel spectrum variations are analyzed using the LPC-cepstrum distance to introduce a quantitative measure for unit selection. Using 5,240 words, vowel spectral distortion resulting from the contextual differences was compared, and the following tendencies were found:(1)The following consonant, the position in the utterance, and the accentuation affect vowel spectral envelopes in above order. (2)For CVs whose following consonants have the same point or manner of articulation, the spectral distance among the vowels of the CVs is 12% smaller than the average. (3)The vowel spectrum of the word peripheral CV differs from that of the word medial CV. Based on these results, a quantitative measure is introduced to represent spectral similarities of each vowel. With this measure, the unit selection scheme was tested using open data. Through these experiments, it was not only confirmed that the previously proposed categorical measures are adequate for general unit selection, but also shown that some phoneme combinations should be specially scored for unit selection.

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

1. INTRODUCTION

Aiming at the efficient use of speech segments in a given speech data set according to an input phoneme string, we have proposed the speech synthesis scheme using non-uniform units. In this system, any speech segments which correspond to phoneme substrings of the given set can be used as unit templates with their contextual information. The selection of a non-uniform unit template sequence is carried out to minimize the speech quality degradation that results from a mismatch between a unit template and the desired context and from the discontinuities between units. The unit selection criteria plays an important role for the optimal choice of the unit sequence.

In this paper, we give the experiment results of vowel spectral difference analysis to evaluate the plausibility of the phonemic category and to obtain a quantitative measure based on acoustic data. Furthermore, some experimental results from investigating the limit of optimal unit selection according to the context information are given.

2. SYNTHESIS SYSTEM

Fig.1 shows the outline of the system.[1] In this synthesis scheme, a Synthesis Unit Entry (SUE) dictionary and the corresponding speech segment file are used instead of a traditional speech unit file. This SUE dictionary is made by sorting and merging all sub-phonemic clusters derived from a given speech data set to facilitate the search for appropriate unit candidates. Using this SUE dictionary, synthesis is carried out in the following manner.

(1) A synthesis unit candidate is obtained for a given phoneme input sequence by searching the SUE dictionary.
(2) Because each entry can correspond to multiple acoustic samples in different utterances, some samples are selected for each entry by comparing the unit extraction context with the target context for synthesis.
(3) In the unit candidate lattice, the most preferable unit combination is selected by assessing the appropriateness of the unit.
(4) Finally, after modifying these speech templates using prosodic information given by rules, conventional synthesis is carried out.

As shown in this synthesis scheme, unit selection criteria play an important role in the optimal choice of the unit sequence.

3. VOWEL SPECTRAL VARIATIONS

To establish a quantitative unit selection measure, a speech database consisting of 5,240 Japanese words spoken by a professional narrator was used.[2] The database is labeled with acoustic-phonetic events. The number of syllables used in computing is shown in Table 1. Using this database, vowel spectral variation resulting from the contextual differences were computed. For these contextual differences, three factors are taken into account. They are (1) following consonant (2) syllable position in the utterance (initial, medial or final), and (3) accentuation.

As the spectral variation measure, we adopted the averaged LPC cepstral distance between two vowel portions in different contexts. (For this calculation 16 LPC cepstral coefficients are calculated every 2.5 msec (in each 21.3 msec Hamming window)). To align the speech segments, linear time scaling is carried out using each vowel duration.

First, the CV unit's vowel samples are clustered according to the contexts for each CV. Using the centroid of the cluster, LPC cepstral distances among clusters were computed. Experiment results are shown in Figures 2, 3, and 4.

Figure 2 compares the ratio of vowel spectral variations resulting from contextual differences to the variance of clusters having the same contexts. As shown in this figure, it was discovered that the following consonant, the position in the utterance, and the accentuation affect vowel spectral envelopes in this order. As the effect of following consonant is greatest, we focused on the effect of following consonants. To compare the variance due to following consonants, LPC cepstral distance is calculated for each vowel cluster followed by the consonant having the same point or manner of articulation.

Figure 3 compares the ratio of vowel spectral variations resulting from following consonant to the variance of clusters for vowels followed by non-specified consonants, (a) among consonants of the same manner of articulation, and (b) among consonants of the same point of articulation. For CVs whose following consonants have the same point or manner of articulation, the spectral distance among the vowels of the CVs is 12% smaller than the average.

Figure 4 compares the ratio of vowel spectral variations resulting from syllable position to the variance of clusters having the same contexts. The vowel spectrum of the word peripheral CV differs from that of the word medial CV.

4. THE EFFECTIVENESS OF CONTEXTUAL INFORMATION FOR UNIT SELECTION

To confirm the appropriateness of the unit selection according to the context, the speech data selected by the target context were compared with the data for each CV centroid using data consisting of 544 CV samples in 310 words. Case A is the same contextual unit selection as the target, and case B is the CV centroid selection. Each unit spectrum selected by each method was compared with the spectrum of the original speech. For all samples, the unit selected by method A is is equal to the unit selected by method B or more similar to the original than that in 95% of the cases.

More precisely, the distribution of difference of vowel spectral distance from the original in cases A and B is shown in Figure 5. From this figure, the following facts can be found. (1) As there are fewer samples whose differences are lower than -0.2 than those whose differences are greater than 0.2 (the value of 0.2 is the averaged vowel spectral variations among vowels having the same contexts), the CV data selected by contextual information is slightly better than the CV centroid data. (2) In some cases, CV selection according to the contextual information is not good.

This result shows that the contextual phonemic similarities are not always sufficient to assign the speech data for selected units. Fine acoustic information should be given to each speech datum to avoid mis-selection.

5. CONCLUSIONS

In this paper, we gave the experiment results of vowel spectral difference analysis and the effectiveness of contextual information for unit selection. From these analyses, we drew the following conclusions.

(1) As the vowel spectral difference is greatly dependent on contextual similarities, contextual similarities can serve as a rough unit selection measure.

(2) In the contextual dependencies, the effect of following consonant is much greater than that of phoneme location and that of accent.

(3) For the measure of unit selection, contextual similarities are not sufficient. It is necessary to incorporate more detail information into the selection measure.

ACKNOWLEDGEMENTS

The author thanks Dr.Kurematsu for his continuous support of this research and also Dr.Shikano for fruitful discussions.

REFERENCES

[1] Y.Sagisaka, "Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units," ICASSP-88 pp.679-682, April, 1988

[2] K.Takeda, Y.Sagisaka, S.Katagiri, "Acoustic-Phonetic Labels in a Japanese Speech Database," Proc. of Euro. Conf. on Sp. Tech. Vol.II pp.13-16,1987

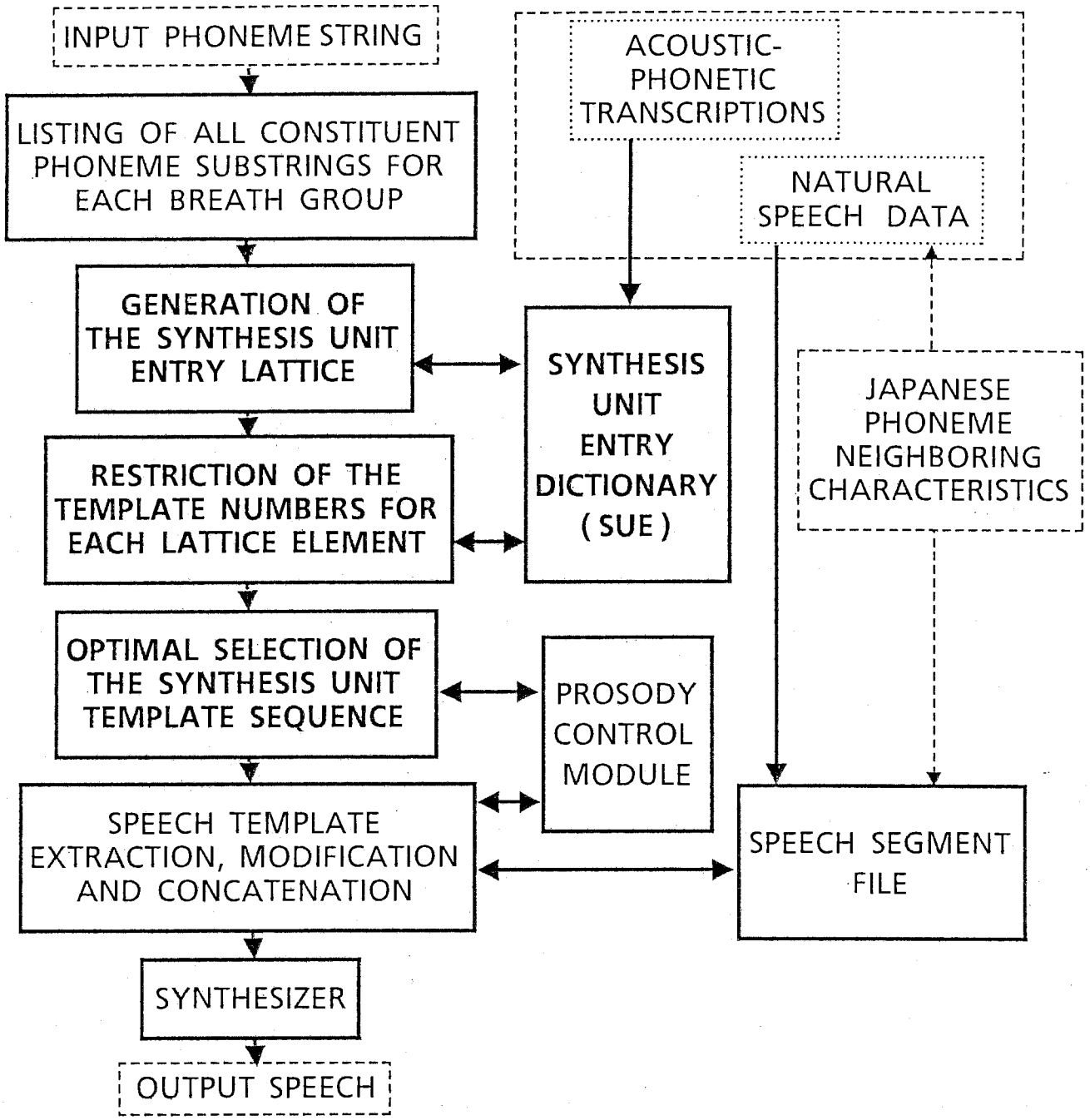

Fig.1 Outline of the Synthesis Scheme Using Non-Uniform Units

Table 1 Number of CV unit samples

Vowel / Consonant	a	i	u	e	o
k	448	316	698	157	221
s	218	235	275	110	115
t	247	90	409	100	241
n	152	78	21	60	94
h	145	68	127	8	48
m	220	166	159	118	145
y	137	-	45	-	75
r	73	168	671	27	52
w	154	-	-	-	-

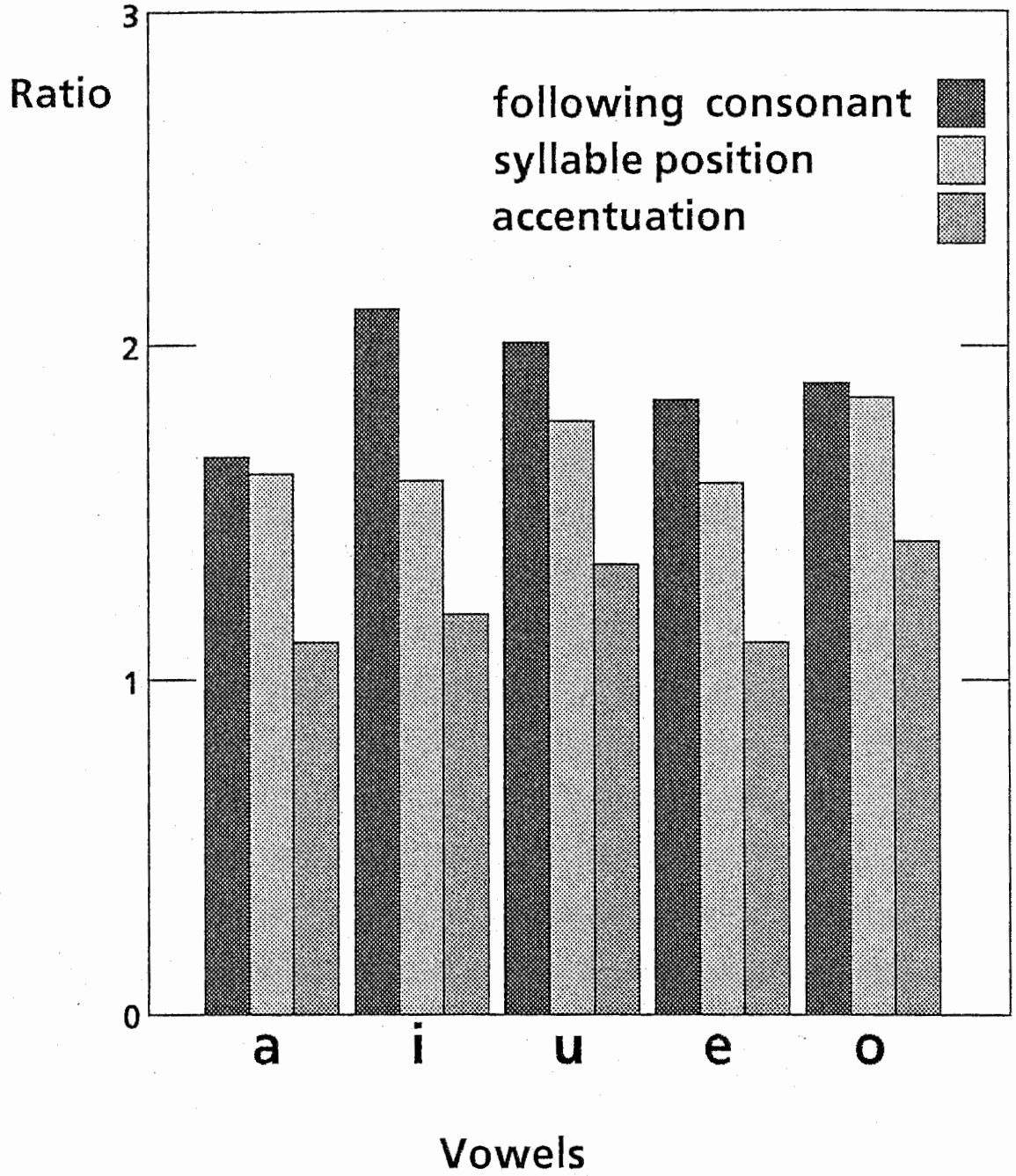

Fig.2 Comparison of the ratio of vowel spectral variations resulting from contextual differences to the variance of clusters having the same context

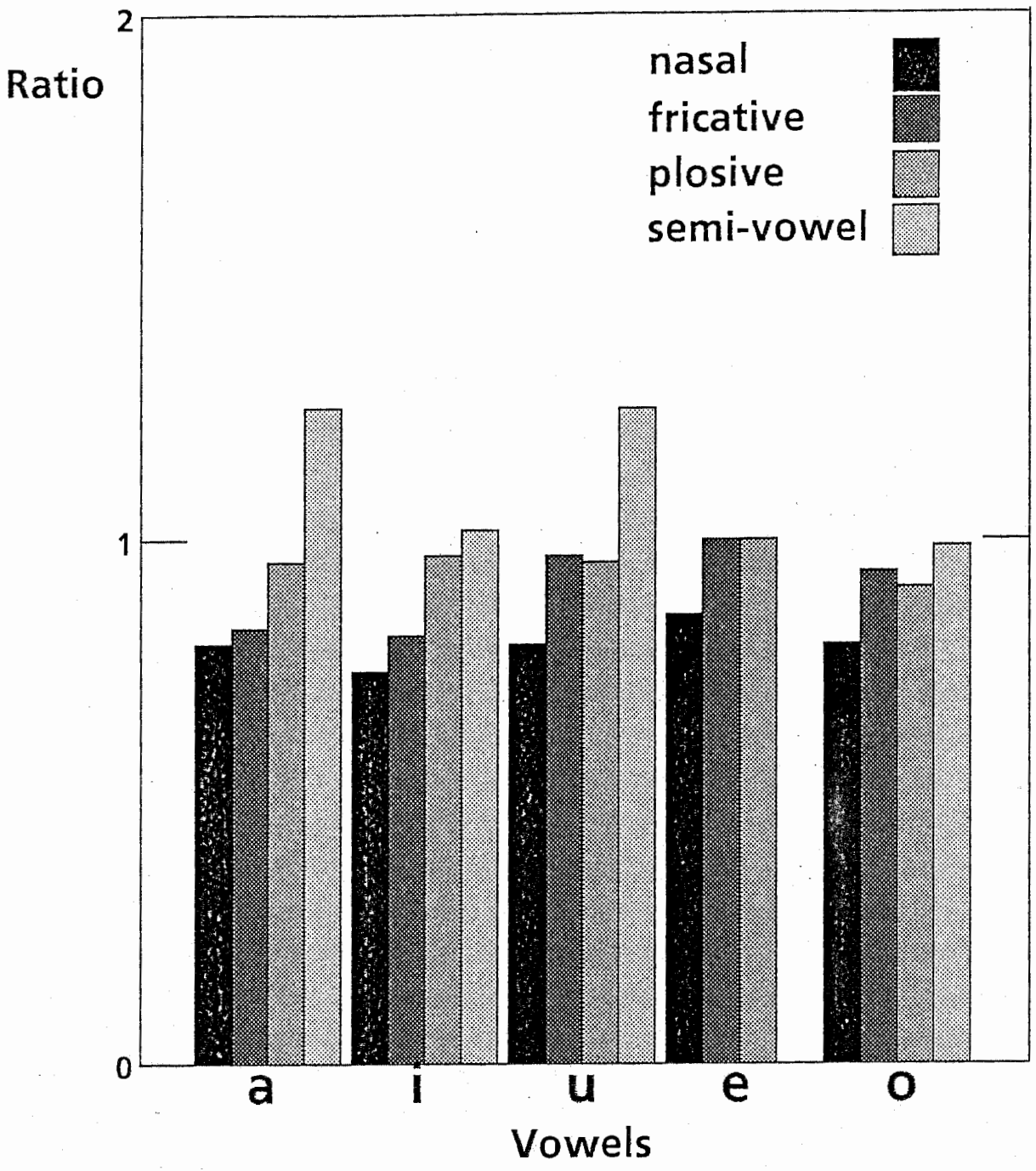

Fig.3 Comparison of the ratio of vowel spectral variations resulting from following consonants to the variance of clusters for vowels followed by non-specified consonants (a) among consonants of the same manner of articulation

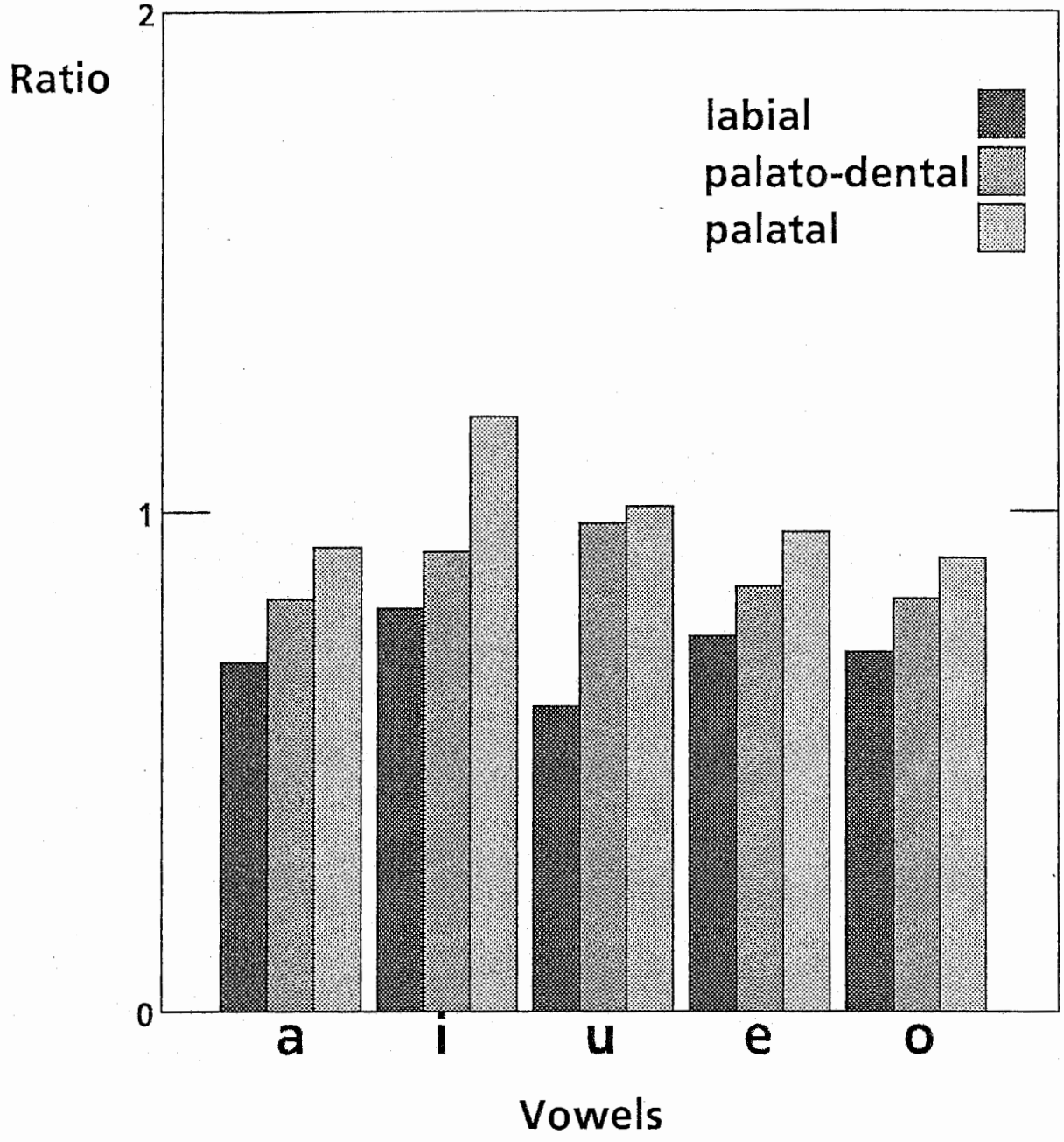

Fig.3 Comparison of the ratio of vowel spectral variations resulting from following consonants to the variance of clusters for vowels followed by non-specified consonants (b) among consonants of the same point of articulation

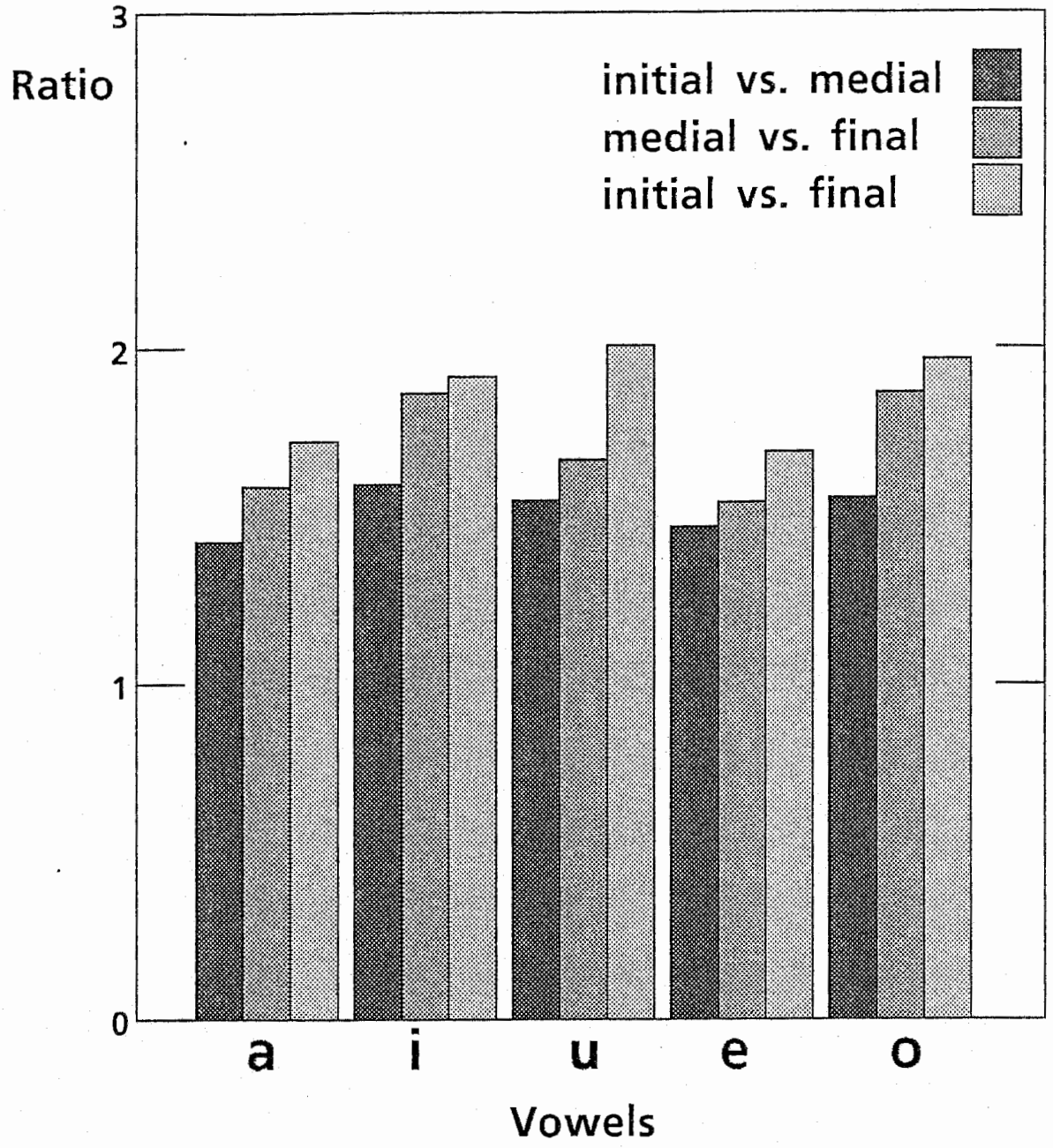

Fig.4 Comparison of the ratio of vowel spectral variations resulting from syllable position to the variance of clusters having the same context

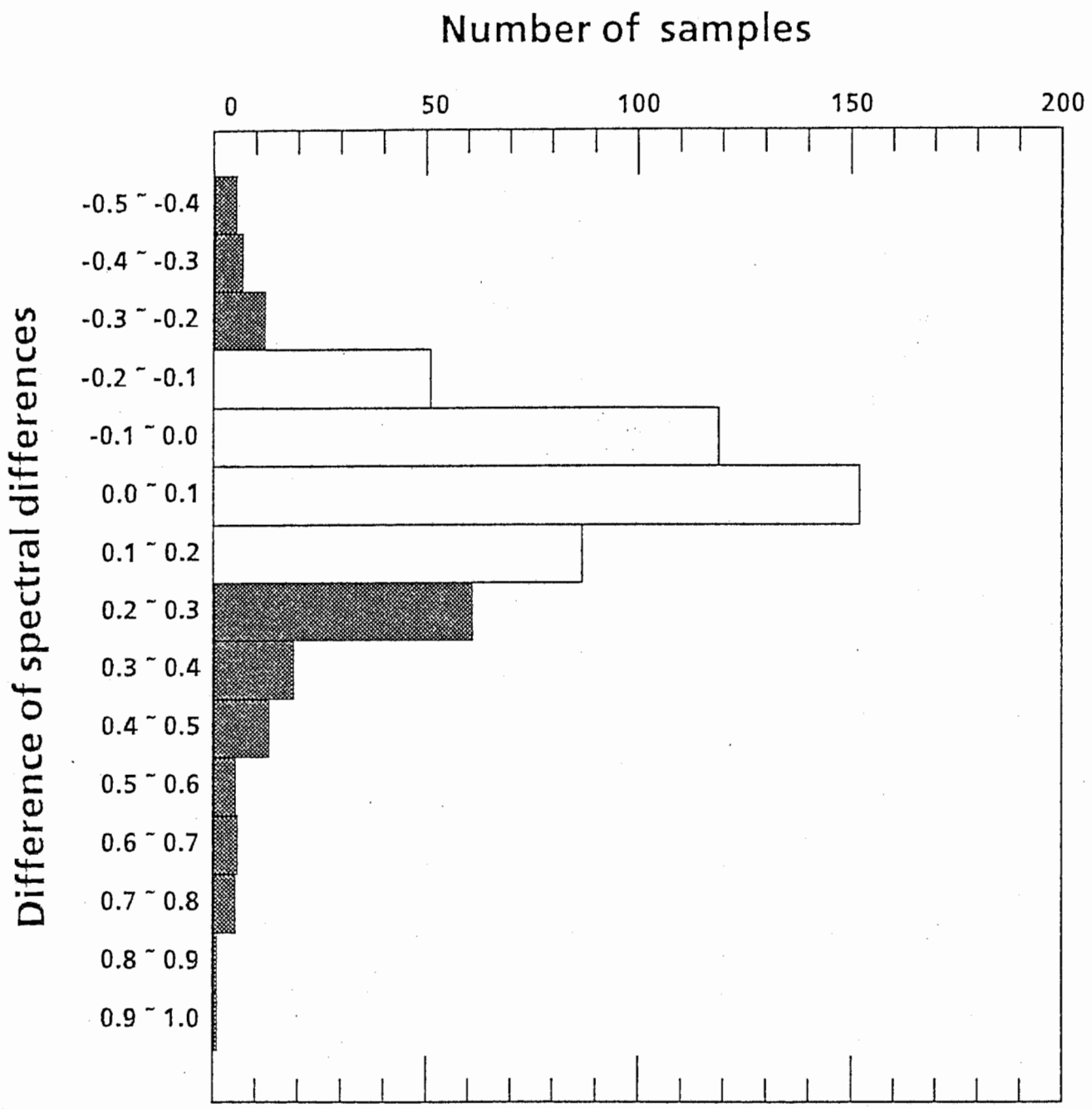

Fig.5 Distribution of differences of vowel spectral distance to the original in cases A and B.