ACOUSTICAL LETTER

Control of phoneme duration based on the movement of speech organs

Norio Higuchi, Tohru Shimizu, Hisashi Kawai, and Seiichi Yamamoto

KDD R & D Laboratories,

2-1-15, Ohara, Kamifukuoka-shi, Saitama, 356 Japan

(Received 22 December 1992)

Keywords: Speech synthesis, Speech production process, Phoneme duration control,
Linear regression model, Quantification theory
PACS number: 43. 72. Ja

1. Introduction

The quantification method type I^{1,2)} has recently often been used for the control of phoneme durations.^{3,4)} In the previous studies there is a fatal problem as regards the way of modelling, because the category of the phoneme whose duration should be determined as the categories of its neighboring phonemes are assumed to be independent affecting factors, which are called items in this paper. According to such an assumption, the categories of contiguous phonemes affect phoneme durations independently and additively. This, however, is not acceptable when we consider the speech production process precisely. The amount of effort required to pronounce phonemes is determined not by the instantaneous positions of the speech organs but by their movements in the transition from the preceding pho-

neme to the following phoneme, and it is thought to be closely related to the phoneme durations.

The authors propose a new method to control phoneme durations based on the movement of speech organs which are used as items in quantification method type I. In order to evaluate its validity, durations of the phonemes included in sentence utterances spoken by two professional male narrators are estimated. The result of the experiment shows that the average estimation errors of the phoneme durations are around twenty milliseconds, and that the total number of categories is much smaller than that of the previous studies.

2. Modelling based on the movement of speech organs

If the categories of contiguous phonemes are assumed to be items, a certain value which is called a score is

Table 1 Items affecting phoneme durations.

(a) For consonants

Item	Number of categories
Manner of speech of controversial consonant	12
Repetitive articulation between preceding consonant and controversial consonant	3
Repetitive articulation between controversial consonant and following consonant	3
Occurance of semivowel /j/ in preceding syllable or controversial syllable	4
Movement of jaw from preceding vowel to following vowel	9
Movement of tongue from preceding vowel to following vowel	5
Gemination of controversial consonant or following consonant	4
Location in linguistic phrase	9

(b) For vowels

Item	Number of categories
Category of vowel (including distinction of elongated or non-elongated vowel)	11
Movement of jaw from preceding vowel to controversial vowel	6
Movement of jaw from controversial vowel to following vowel	6
Movement of tongue between controversial vowel and adjacent vowels	8
Manner of speech of preceding consonant	10
Manner of speech of following consonant	10
Occurance of semivowel /j/ in controversial syllable or following syllable	4
Repetitive articulation between preceding consonant and following consonant	3
Accentuation	3
Location in linguistic phrase	9

always added to or subtracted from the standard phoneme duration when preceding phoneme is /i/ without any consideration of the category of the following phoneme. The influence of the phenomena wherein the preceding phoneme is a narrow vowel /i/ is considered to be different in the case that following vowel is a wide vowel and in the case that it is a narrow vowel.

To avoid this problem the movement of speech organs are used for the control of phoneme durations. Furthermore, since a syllable is a basic unit in the pronunciation of Japanese, and consonants and vowels are intrinsically different from each other, the interaction of two consonants surrounding a vowel and that of two vowels surrounding a consonant sequence are also taken into consideration for modelling phoneme duration control. Table 1 (a) and (b) show the items for modelling consonant durations and vowel durations respectively.

3. Speech materials

The speech materials used in the evaluation test of the model are 503 sentence utterances spoken by a male speaker A (MYI) recorded at ATR⁵⁾ and 250 sentence utterances spoken by a male speaker B recorded at KDD. 250 sentences are A, B, C, E and F sets of

Table 2 Number of phonemes used in evaluation test.

Speaker	Consonant	Vowel
Speaker A	10,624	12,384
Speaker B	5,782	6,624

ATR sentences and they are a part of 503 sentences.

For the evaluation test, the durations of the phonemes of which both ends can be determined by manual segmentation based on the rapid change of the spectrum are used. Table 2 shows the numbers of consonants and vowels used in the evaluation test. The segmentation is performed according to the definition of phonemes which is used for speech synthesis systems by authors.⁶⁾ According to the above-mentioned definition of phonemes, a semivowel /j/ occurs only in /mjV/-, /njV/-, /rjV/-, /bjV/- and /pjV/-syllables, where V indicates vowels /a/, /i/, /u/, /e/ or /o/. Other syllables are divided into a palatalized consonant and a vowel, as for example /kja/ is divided into [c] and [a].

4. Result of the evaluation test

Table 3 (a) and (b) show the partial correlation

Table 3 Partial and multiple correlation coefficients.

(a) For consonants

Item	Speaker A	Speaker B
Manner of speech of controversial consonant	0.666	0.714
Repetitive articulation between preceding consonant and controversial consonant	0.028	0.044
Repetitive articulation between controversial consonant and following consonant	0.071	0.018
Occurrence of semivowel/j/ in preceding syllable or controversial syllable	0.201	0.044
Movement of jaw from preceding vowel to following vowel	0.121	0.221
Movement of tongue from preceding vowel to following vowel	0.100	0.065
Gemination of controversial consonant or following consonant	0.667	0.562
Location in linguistic phrase	0.470	0.485
Multiple correlation coefficient	0.837	0.835

(b) For vowels

Item	Speaker A	Speaker B
Category of vowel (including distinction of elongated or non-elongated vowel)	0.760	0.741
Movement of jaw from preceding vowel to controversial vowel	0.107	0.061
Movement of jaw from controversial vowel to following vowel	0.229	0.119
Movement of tongue between controversial vowel and adjacent vowels	0.510	0.602
Manner of speech of preceding consonant	0.503	0.519
Manner of speech of following consonant	0.403	0.252
Occurrence of semivowel /j/ in controversial syllable or following syllable	0.117	0.065
Repetitive articulation between preceding consonant and following consonant	0.025	0.025
Accentuation	0.084	0.105
Location in linguistic phrase	0.486	0.606
Multiple correlation coefficient	0.822	0.827

Table 4 Standard error of duration.

Speaker	Consonant	Vowel
Speaker A	19.23 ms	16.88 ms
Speaker B	20.92 ms	19.47 ms

coefficients of each item and the multiple correlation coefficients of the estimation of consonant and vowel durations respectively. In the control of consonant durations, the manner of speech of the controversial consonant, the gemination of the controversial consonant or the following consonant and their location in the linguistic phrase are the dominant factors. On the other hand, dominant factors in the control of the vowel durations are the category of the vowel including the distinction as to whether or not the vowel is elongated, the movement of the tongue from the controversial vowel to the following vowel, the manner of speech of the preceding consonant, the position in the linguistic phrase and the manner of speech of the following consonant. This result shows that the influence of the articulation of consonants on the vowel durations is large, though the converse effect is small. It also shows the intrinsic difference between consonants and vowels.

In quantification method type I, each score is expected to have some meaning. For example, a repetitive articulation of the same consonant takes a longer time than other consonant-consonant combinations, because it needs a rapid change in the direction of movement of speech organs. The result of the evaluation test supports this hypothesis. In the case of the repetition of the lip articulation, the first consonant and the intermediate vowel are together ten milliseconds longer than those in other combinations, and the intermediate vowel and the second consonant are twenty milliseconds longer in the case of the repetitive tongue articulation.

The average errors of the estimated phoneme durations are shown in Table 4.

5. Conclusions

The authors proposed a new duration control model which is based on the movement of speech organs in addition to the consideration of the difference between consonants and vowels. This model can estimate phoneme durations using eight and ten items, and forty-nine and seventy categories for consonants and vowels respectively. The average error of the estimated durations is shorter than or almost equal to twenty milliseconds.

References

- 1) C. Hayashi, "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view," *Ann. Inst. Stat. Math.* **3**, 69–98 (1952).
- 2) C. Hayashi, "New developments in multidimensional data analysis," in *Recent Developments in Clustering and Data Analysis*, E. Diday, C. Hayashi, M. Jambu, and N. Ohsumi, Eds. (Academic Press Inc., Boston, 1988), pp. 3–16.
- 3) T. Sakayori, S. Sasabe, and H. Kitagawa, "Prosody control using categorical multiple regression for speech synthesis by rule," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 245–246 (1986) (in Japanese).
- 4) K. Takeda, Y. Sagisaka, and H. Kuwabara, "On sentence-level factors governing segmental duration in Japanese," *J. Acoust. Soc. Am.* **86**, 2081–2087 (1989).
- 5) M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database user's manual," ATR Tech. Rep. TR-I-0166 (1990) (in Japanese).
- 6) N. Higuchi, S. Yamamoto, and T. Shimizu, "A Japanese speech synthesizer based on production rules," *J. Acoust. Soc. Am.* **84** (Suppl. 1), J18 (1989).