SEGMENTAL DURATION CONTROL BASED ON AN ARTICULATORY MODEL

Yoshinori Shiga, Hiroshi Matsuura and Tsuneo Nitta

Multimedia Engineering Laboratory, TOSHIBA Corporation
70 Yanagi-cho, Saiwai-ku, Kawasaki, Japan

ABSTRACT

This paper proposes a new method that determines segmental duration for text-to-speech conversion based on the movement of articulatory organs which compose an articulatory model.

The articulatory model comprises four time-variable articulatory parameters representing the conditions of articulatory organs whose physical restriction seems to significantly influence the segmental duration. The parameters are controlled according to an input sequence of phonetic symbols, following which segmental duration is determined based on the variation of the articulatory parameters.

The proposed method is evaluated through an experiment using a Japanese speech database that consists of 150 phonetically balanced sentences. The results indicate that the mean square error of predicted segmental duration is approximately 15[ms] for the closed set and 15-17[ms] for the open set. The error is within 20[ms], the level of acceptability for distortion of segmental duration without loss of naturalness, and hence the method is proved to effectively predict segmental duration.

1. INTRODUCTION

Duration control is one of the most important factors that decide the naturalness of speech produced by text-to-speech(TTS) systems. Unnatural and artificial duration makes speech so monotonous that people quickly become tired of listening to it, and often causes misperception.

In actual speech production, as illustrated in Figure 1, plans are first made on local speaking rates according to prosodic information, such as word stress, syntactic structure and semantic focuses, i.e., the target timing structure is determined. Although voice is then adjusted to the target structure, articulatory organs can not always maintain the voice at the target due to physical restrictions on movement of the organs. We focus on this phenomenon and believe that it is the main factor affecting the basic timing structure of each language and also prevents synthetic speech from being monotonous. Therefore, in order to synthesize temporally natural-sounding speech, segmental duration should be controlled with consideration of the restricted movement of organs.

In this paper, we propose a novel method for determining segmental duration based on the movement of articulatory organs which compose an articulatory model, in order to take into account the movement restriction of articulatory organs. The method has already been applied to duration control in our Japanese TTS system that runs on PCs, and helps to make synthetic speech much more natural.

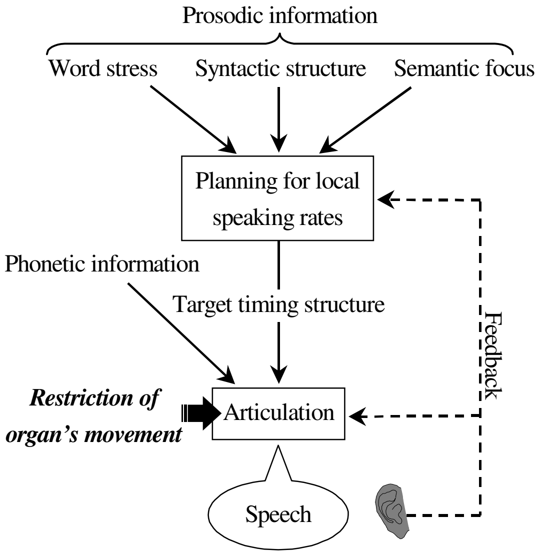

Figure 1: Duration control in actual speech production.

2. DURATION CONTROL BASED ON AN ARTICULATORY MODEL

2.1 Articulatory Model

Many kinds of articulatory model have been proposed[1-3] after the earliest one proposed by Coker and Fujimura[4]. Most of these models are, however, designed to approximate the transmission characteristics of the vocal tract in order to clarify the relation between the movement of vocal organs and acoustic characteristics of speech. Consequently, they are complicated by having many parameters corresponding to every part of the vocal tract.

On the other hand, the articulatory model in our method for determining segmental duration is fairly simple, because it is sufficient for the model to employ parameters that represent the conditions of articulatory organs whose physical restriction significantly influences the segmental duration. Figure 2 shows the articulatory model we adopt in the proposed method. The

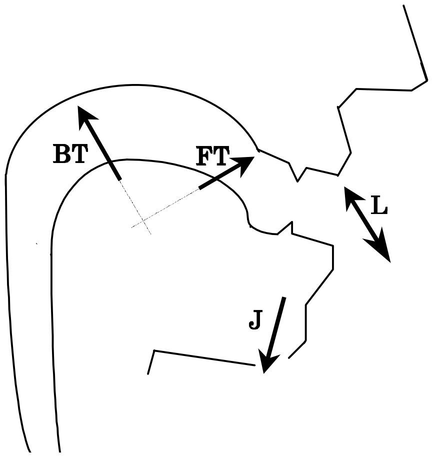

Figure 2: The articulatory model in the proposed method.

model comprises only four time-variable articulatory parameters, i.e., the opening area of the lips (L), and the positions of the lower jaw (J), the front tongue (FT) and the back tongue (BT), whose physical restriction is considered to significantly influence the duration.

2.2 Model Control

Since the method simulates the movements of the articulatory organs, speech sounds should be classified according to their manners and places of articulation. The classification is done using a knowledge of articulatory phonetics. We will discuss the classification of Japanese speech in section 3.1.

There are 13 coefficients assigned to each classified sound (simply called "phone" below) in total. They include three coefficients of each articulatory parameter (therefore 12 values in total), $\boldsymbol{A}_{inh}$, $\boldsymbol{A}_{max}$ and $\boldsymbol{A}_{min}$, whose values represent the inherent articulation and the upper and lower articulatory limits of the phone, respectively. The coefficient $\boldsymbol{A}_{inh}$ indicates the organ's "typical" position and $\boldsymbol{A}_{max}$ and $\boldsymbol{A}_{min}$ the range of the position acceptable as the articulation of the phone. The remaining coefficient is the minimum duration $\boldsymbol{D}_{min}$ of the phone; this duration must be put after every parameter comes into articulation of the phone until a command for articulation of the next phone is given.

Based on these coefficients, the articulatory parameters are controlled in the model according to an input sequence of phonetic symbols. We approximate the parameter change of the articulatory organ represented by $\boldsymbol{k}$(= L,J,FT,BT) with the following function ***M(k,t)***:

$$M(k,t) = A_{inh}(k, ph_0) + \sum_{i=1}^{N-1} R_i(k,t)$$

$$R_i(k,t) = \begin{cases} 0 & (t < t_i) \\ \{A_{inh}(k, ph_{i+1}) - A_{inh}(k, ph_i)\} S(t - t_i) & (t \geq t_i) \end{cases}$$

where $\boldsymbol{ph_i}$ and $\boldsymbol{N}$ indicate the type of the $\boldsymbol{i}$-th phone and the number of phones to be synthesized, respectively. Here, $\boldsymbol{t_i}$ is the time where all the articulatory parameters start shifting the articulation from the $\boldsymbol{i}$-th phone to the next phone. We suppose that a "command" for the phone $\boldsymbol{ph_{i+1}}$ is given at the time $\boldsymbol{t_i}$.

S(t) is obtained by the following step-response function of a critically damped second-order system:

$$S(t) = 1 - (1 + \alpha_k t)e^{-\alpha_k t}$$

where $\boldsymbol{\alpha_k}$ is a time constant inherent in the articulatory organ $\boldsymbol{k}$. An organ that moves quicker takes a higher value of $\boldsymbol{\alpha}$.

Figure 3 shows the parameter change representing the spring-up and –down movements of the tongue, which are realized with two commands. Figure 4 shows an example of articulatory parameter variation that the method generates from the input "*arayuru genjitsu* (all the facts)".

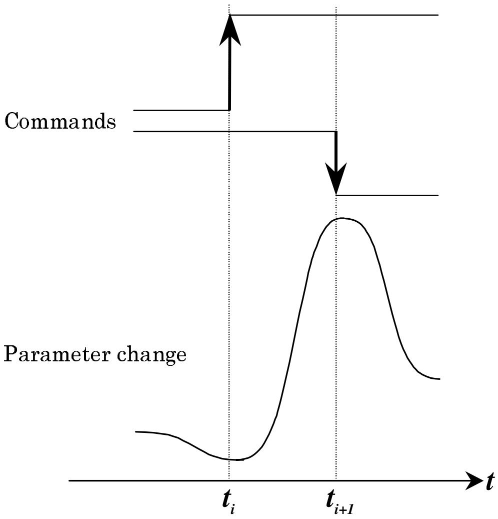

Figure 3: Commands and the change of articulatory parameter.

2.3 Duration Determination

Segmental duration is determined based on the four contours of the time-variable articulatory parameters, which are controlled according to the input phonetic string to be synthesized. The method first sets the boundaries of segments. Although there are generally different definitions of segment boundary, the definition must be at least the same as that in the post-processing. Our TTS system employs a concatenative method using diphone units for post-processing, which are produced on the basis of phonetic labels in the speech database we built. The phonetic boundaries are hence set corresponding to the definition of the labeling method of the database, then segmental duration is obtained as the time difference between adjacent boundaries.

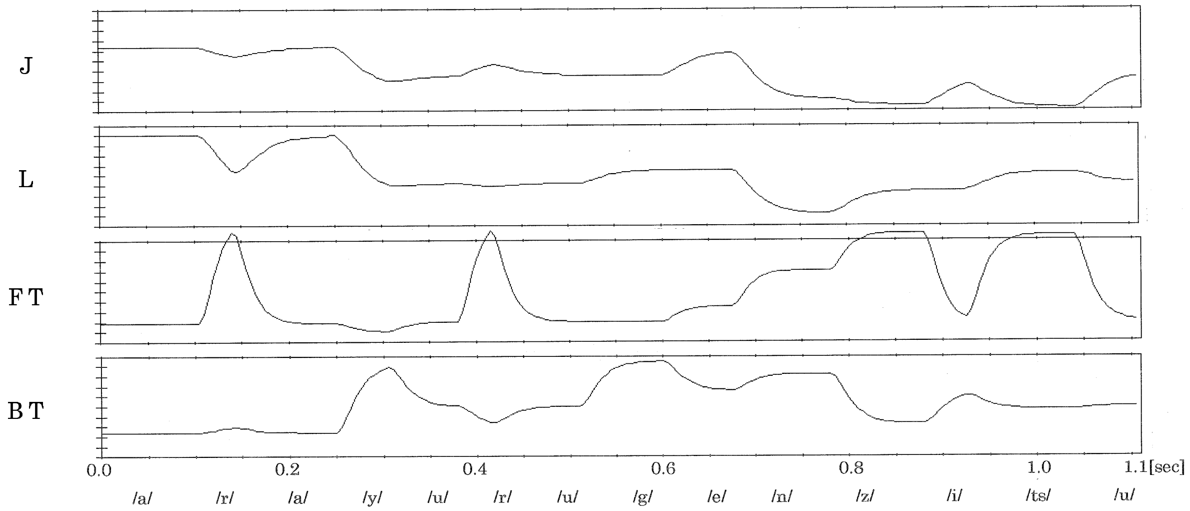

Figure 4: Articulatory parameter variation that the method generates from the input “*arayuru genjitsu* (all the facts)”.

3. APPLICATION TO JAPANESE SPEECH SYNTHESIS

We have discussed the proposed method independent of language so far. In this section, the method is applied to Japanese speech.

3.1 Classifying the Sound

As explained in section 2.2, speech sounds must be classified by their manners and places of articulation. In order to meet this requirement, we referred to the classification of International Phonetic Alphabet (IPA). With this reference, we classified Japanese sounds into 51 phones, which consist of 12 vowels including devoiced or nasalized vowels, and 39 consonants including palatalized sounds, which are listed in Table 1. The coefficients for the sound [i] of Japanese are given in Table 2 as an example.

3.2 Controlling the Articulatory Model Based on Mora-timed Rhythm

The timing structure of Japanese speech is characterized by mora-timed rhythm, i.e., rhythm on mora isochrony. This timing structure is achieved with the articulatory model by adjusting each command in a vowel for the following phone so that the commands are issued at equal time intervals. However, if not every articulatory parameter has reached the acceptable range for that vowel, the command waits until this state is reached, because it indicates that the articulatory combination is too difficult to be produced within the given interval of time. A command in a consonant for the next phone is issued after the elapse of the minimum duration D_{min} of the consonant from where every articulatory parameter comes into its acceptable articulation range, as explained in section 2.2.

Vowels	a i ɯ e o i̥ ɯ̥ ã ĩ ɯ̃ ẽ õ
Consonants	k k^j s ɕ ʃ t t^j ts tɕ tʃ n n^j ɲ h ç ɸ ɸ^j m m^j j ɾ ɾ^j w p p^j ɡ ɡ^j z dz dʑ dʒ d d^j b b^j N ʔ ŋ ŋ^j

Table 1: Sounds in Japanese speech.

A_{min}(J, i) A_{inh}(J, i) A_{max}(J, i)	0.19 0.30 0.50
A_{min}(FT, i) A_{inh}(FT, i) A_{max}(FT, i)	0.10 0.17 0.25
A_{min}(BT, i) A_{inh}(BT, i) A_{max}(BT, i)	0.59 0.64 0.83
A_{min}(L, i) A_{inh}(L, i) A_{max}(L, i)	0.32 0.35 0.49
D_{min}(i)	0.00

Table 2: Example of the coefficients (Japanese [i]).

4. EVALUATION

We objectively evaluate the proposed method by examining errors between actual duration and the duration predicted by the method.

4.1 Speech Data

The data used in the experiment are 150 phonetically-balanced sentences in Japanese. Labels representing the phone types that we classified in section 3.1 are manually assigned. 100 sentences out of all the data are used as a closed set and the remaining 50 as an open set.

4.2 Experimental Procedure

The coefficients of each phone are first assigned values estimated roughly from the articulatory shape with a knowledge of articulatory phonetics, and then optimized by the A-b-S method with the closed set of 100 sentences.

As the interval of time until a command in a vowel for the next phone is issued, we use the value extracted from the database by the "accent phrase", assuming that the interval is constant in the phrase.

Segmental duration is predicted for all the sentences using the proposed method, and compared to the measured duration for the same 100 sentence data used for the optimization, and for the remaining 50 sentence data for an open test.

4.3 Results

The experimental results are shown in Table 3. The mean square error was approximately 15[ms] for the closed set and 15-17[ms] for the open set.

	Closed(100sentences)		**Open(50sentences)**	
	Consonant	**Vowel**	**Consonant**	**Vowel**
Duration average[ms]	60.5	79.8	61.7	77.7
Mean square error[ms]	14.9	15.4	15.4	17.3

Table 3: Experimental Results (Mean square error between estimated duration and actual duration).

5. DISCUSSION

In order to evaluate the method from experimental results, the acceptability for distortion of segmental duration should be taken into consideration.

Hashimoto[5] reported that approximately 20[ms] is the limit of segmental duration within which speech remains natural. All the mean square errors in Table 3 are within the limit, and hence the experimental results show that the proposed method precisely estimates segmental duration.

6. CONCLUSION

We have discussed a new method that determines segmental duration by simulating articulatory motion with an articulatory model. After a theoretical examination, an experiment was conducted, and the results confirmed the effectiveness of the proposed method.

Since the method determines duration based on the articulatory model representing the movement of articulatory organs, duration is influenced by physical restriction of the model fairly similar to the process of actual speech production. The proposed method is therefore expected to give a natural rhythm to synthetic speech at different speaking rates, but not to be good at tongue twisters.

ACKNOWLEDGMENT

We are very grateful to the members of the speech synthesis group in the TOSHIBA R&D Center for willingly providing valuable speech databases to us.

REFERENCES

1. Mermelstein, P. "Articulatory model for the study of speech production," *J. Acoust. Soc. Am. 53*(40): 1070-1082, 1973.
2. Shirai, K. and Honda, M. "Estimation of articulatory motion from speech waves and its application for automatic recognition," in Spoken Language Generation and Understanding (ed. J. C. Simon), Reidel, Dordrecht, Holland, pp.87-99, 1980.
3. Maeda, S. "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modeling* (NATO Advanced Study Institute Series), W. J. Hardcastle and A. Marchal (eds). Kluwer Academic Publishers, Boston, pp.131-149, 1990.
4. Coker, C. H. and Fujimura, O. "A model for specification of vocal tract area function," *J. Acoust. Soc. Am. 40*: 1271(A), 1966.
5. Hashimoto, S. and Saito, S. "Prosodic rules for speech synthesis," 7th International Congress on Acoustics, pp.129-132, 1971.