論文 / 著書情報
Article / Book Information

Title	Speech-Rate-Variable HMM-Based Japanese TTS System
Author	Koji Iwano, Masahiro Yamada, Taro Togawa, Sadaoki Furui
Journal/Book name	IEEE 2002 Workshop on Speech Synthesis (TTS 2002), Vol. , No. , pp.
Issue date	2002, 9
DOI	http://dx.doi.org/10.1109/WSS.2002.1224413
URL	http://www.ieee.org/index.html
Copyright	(c)2002 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

SPEECH-RATE-VARIABLE HMM-BASED JAPANESE TTS SYSTEM

Koji Iwano, Masahiro Yamada, Taro Togawa, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
{iwano,masahiro,tarot,furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes a new method for controlling phoneme duration according to arbitrary target speech rate in speech synthesis (TTS, text-to-speech) systems. The proposed method first constructs three fundamental duration models at "fast", "normal", and "slow" speech rates using Hayashi's Quantification Theory (Type 1) based on real speech databases and creates a duration model according to a target speech rate by interpolating the fundamental models. Our TTS system uses an HMM-based synthesizer which can achieve flexible prosody control. Various speech synthesized by the proposed method are evaluated by subjective experiments at four speech rates using pair comparison tests between the proposed method and a rule-based method. The results show that the proposed method achieves higher naturalness in synthesized speech than the rule-based method.

1. INTRODUCTION

Quality of synthesized speech has recently improved significantly and text-to-speech conversion systems (TTS systems) are now used for many applications. However, most of them can only produce reading-style speech and there is a strong demand for a more flexible TTS systems that can produce a variety of speech styles to be used in a wider scope of applications.

Prosody control is essential for achieving various styles in speech synthesis. Speech rate control is one of the most important and useful components in the prosody control. For example, generating speech at slow speech rate for aged people and synthesizing speech at fast speech rate for expressing emotions, such as "happiness" or "anger" will become important in the near future. This paper proposes a new method of phoneme duration control according to a given speech rate for synthesizing natural speech at various speech rates.

Most of the present speech synthesizers are based on sub-word unit concatenation methods and they usually use the TD-PSOLA technique [1] for prosody control. Although the TD-PSOLA is a simple and effective method for modifying the prosody, including fundamental frequency (F_0) contours and phoneme duration, it is vulnerable to spectral and phase distortion. On the other hand, synthesizers based on a phonetic vocoder method can control prosody without increasing distortion.

For this reason, our TTS system is based on a phonetic vocoder method using HMM [2, 3]. In this method, phoneme HMMs are used to produce a time function of cepstral parameters. Phoneme duration is modeled and controlled using a statistical method, specifically a categorical multiple regression method called Hayashi's Quantification Theory (Type 1) [4]". For the statistical modeling, we create a speech database consisting of the utterances spoken at three kinds of speech rate: "fast", "normal", and "slow". A phoneme duration model for each speech rate is trained by the categorical multiple regression method and a duration model for an arbitrary speech rate is created by interpolating the three models.

In the first part of the paper, we investigate the most important factors for high-quality phoneme duration modeling based on the Quantification Theory (Type 1) and determine the input factors for duration control. The latter part of the paper describes our phoneme duration control method and results of subjective experiments on the naturalness of the synthesized speech.

2. OUTLINE OF HMM-BASED TTS SYSTEM

Outline of our HMM-based Japanese TTS systems is shown in Figure 1. Japanese text, written with a mixture of Chinese and Kana characters, is given to the system and processed by a text analyzer from the "FLUET" library provided by NTT Cyber Space Laboratories [5]. In the text analysis process, the Japanese text is segmented into "prosodic word [6]" units. The prosodic word unit is a basic Japanese unit with one accent component, corresponding to a word or a word chunk with one or multiple content word(s) and function word(s). For each prosodic unit, the analyzer predicts a phoneme sequence, an "accent type", and a "pitch connection pattern [5]" from the current to the following unit. Each prosodic unit can be modeled by one of the basic accent types. In the Tokyo dialect of Japanese, an n-mora prosodic word is uttered with one of the $n+1$ possible accent types denoted as type i ($i = 0, \cdots, n$) accent [6]. The type 0 accent has no apparent downfall. The other type i accents have rapid downfall in the F_0 contour at the end of the ith mora (syllable). Each prosodic unit boundary has a pitch connection pattern which represents the strength of prosodic connection indicated by one of the six levels; the weakest connection being assigned to long pauses.

In the prosody generation part, both phoneme duration and F_0 contour of the target sentence are determined based on the sequences of the accent types and the phonemes, using the Quantification Theory (Type 1). The details are described in section 3.

The triphone HMMs in Figure 1 are trained using the 50 dimensional feature vectors consisting of 25 mel-cepstral coefficients [7] including the zero-th coefficient and their delta coefficients, extracted from speech signal using a 25.6ms-length Hamming window shifted at every 5ms. Each triphone HMM has five states and four mixtures in each state. The total number of states is approximately 3,000. In the synthesis process, a sentence HMM is constructed by concatenating the triphone HMMs according to the phoneme sequence. A mel-cepstrum vector sequence is obtained from the concatenated HMMs based on the maximum likelihood criterion [8].

The cepstral sequence is then converted into a sequence of MLSA (Mel Log Spectral Approximation) filters [9] and the target

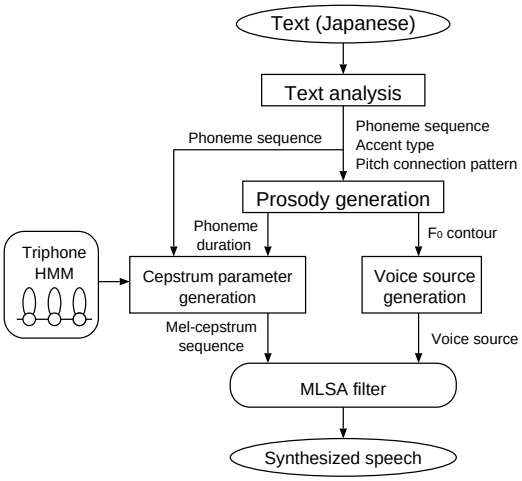

Fig. 1. HMM-based text-to-speech system.

speech is synthesized by passing a voice source waveform through the filters. Pulses and white noises are used as the voice source waveform in voiced and unvoiced parts, respectively.

3. PROSODY GENERATION USING THE QUANTIFICATION THEORY (TYPE 1)

In order to achieve high-quality prosody control in Japanese speech synthesis, several F_0 contour control methods [10, 11, 12] and duration control methods [10, 13] using the Quantification Theory (Type 1) [4] have already been proposed. Our synthesis method uses statistical prosody control based on the quantification theory.

3.1. Quantification Theory (Type 1)

The Quantification Theory (Type 1) formulates the relationship between categorical and numerical values as follows:

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc}\delta_{fc}(i) \qquad (i = 1, \cdots, N) \tag{1}$$

where $\hat{y}_i$ is the predicted value of the ith sample, $\bar{y}$ is the mean value over all samples, and N is the total number of samples. $\delta_{fc}(i)$ is the characteristic function:

$$\delta_{fc}(i) = \begin{cases} 1 & : \text{ if the } i\text{th sample belongs to} \\ & \quad \text{the category } c \text{ of the factor } f \\ 0 & : \text{ otherwise} \end{cases} \tag{2}$$

x_{fc} is the score of the factor f in the category c, which can be computed by minimizing the summation of squared errors $\sum_i (\hat{y}_i - y_i)^2$ using a conventional linear regression model.

3.2. F_0 contour generation

In the prosody generation stage, the F_0 values are predicted, mora by mora, based on the Quantification Theory (Type 1) [14]. The F_0 contour for a sentence is generated by linearly interpolating the moraic values.

In [14], we investigated important factors for F_0 control using the quantification theory. As a result of prediction error estimation and subjective experiments, it has been confirmed that the best F_0 contour is obtained when 21 factors are taken into consideration. These 21 factors are related to the phoneme sequence, the accent types of prosodic units, and the pitch connection patterns.

Table 1. Factors for generating phoneme duration using the Quantification Theory (Type 1). The number in () indicates the number of categories of each factor.

No.	factors
1	Number of mora in W_k (9)
2/3	Number of mora before / after W_k within P (9)
4	Accent type of prosodic unit W_k (7)
5/6	Accent type of prosodic unit W_{k-1} / W_{k+1} (7)
7	Number of prosodic units with accent types higher than 1 before W_k within P (4)
8	Pitch connection pattern at the boundary between W_{k-1} and W_k (4)
9	Pitch connection pattern at the boundary between W_k and W_{k+1} (4)
10	Pitch connection pattern at the boundary between W_{k-2} and W_{k-1} (5)
11	Pitch connection pattern at the boundary between W_{k+1} and W_{k+2} (5)
12	Pitch connection pattern at the boundary between W_{k-3} and W_{k-2} (5)
13	Pitch connection pattern at the boundary between W_{k+2} and W_{k+3} (5)
14/15	Pause length before / after W_k (9)
16	Kind of phoneme O_i (1∼9 : depending on the phoneme class of O_i)
17/18	Kind of phoneme O_{i-1} / O_{i+1} (18∼29 : depending on the kind of phoneme O_i)
19/20	Kind of phoneme O_{i-2} / O_{i+2} (18∼29 : depending on the kind of phoneme O_i)
21	j (9)

3.3. Phoneme duration generation

Phonemes are clustered into 13 classes based on the place of articulation and class-dependent phoneme duration models are constructed. In order to train the duration models, phoneme boundaries in the training corpus are estimated by the Viterbi alignment technique, using the HMMs for the synthesis filter. This means that both the duration models and the HMMs are trained using common units produced by the maximum likelihood criterion. This is very important for avoiding the deterioration of the synthesized speech quality due to mismatches between the units during analysis and synthesis.

Effective factors for predicting phoneme duration are selected by the estimation of prediction error and the results of subjective experiments. They are selected from 21 factors that are similar to those described in section 3.2. Table 1 shows the 21 factors, where the ith phoneme in the sentence is the estimation target phoneme O_i and belongs to the mora M_j. M_j is the jth mora in the prosodic unit W_k. W_k is the kth unit in the sentence and it belongs to the intonational phrase (breath group) P.

A database of 503 phonetically balanced sentence utterances from a male speaker "MHT" which is a part of the ATR continuous speech corpus are used in our experiments. The triphone HMMs for creating the MLSA filter are trained using the whole 503 utterances. On the other hand, the phoneme duration model is trained

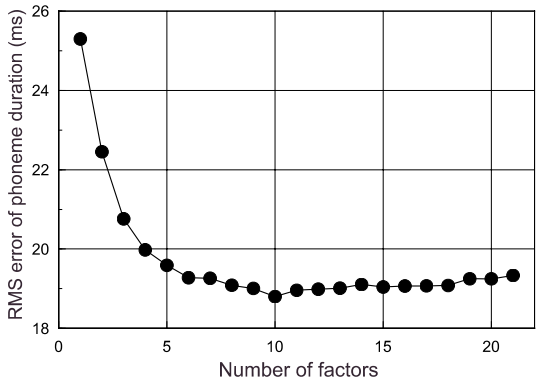

Fig. 2. Relationship between the number of factors selected by the greedy algorithm and the average RMS estimation error of phoneme duration.

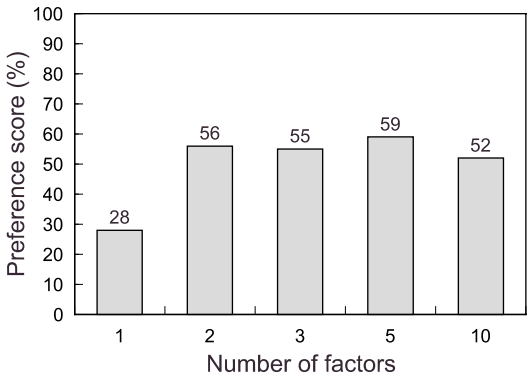

Fig. 3. Preference scores of synthesized speech as a function of the number of factors in the duration control method selected by the Quantification Theory (Type 1).

using 493 utterances and evaluated using the remaining 10 utterances.

The duration model training is performed as follows. First, the "greedy algorithm" is used to select important factors based on the root-mean-square (RMS) estimation error:

1. set $f_n \in C$, $F = \phi$.
   where f_n indicates one of the 21 factors ($n = 1, \cdots, 21$), C is the initial set of factors, and F is the set of selected factors.
2. select f_n from C which minimizes the estimation error when the phoneme duration model is trained using $F + \{f_n\}$.
3. remove f_n from C and add f_n to F.
4. go back to 2 until $C = \phi$.

As the result of the experiment, the following order of importance has been obtained: `18`, `17`, `20`, `16`, `19`, `21`, `15`, `14`, `1`, `8`, `4`, `3`, `5`, `2`, `9`, `6`, `13`, `10`, `12`, `11`, `7`. Figure 2 shows the relationship between the number of most important factors used in the modeling and the averaged RMS estimation error of the phoneme duration. The best result is obtained when the top 10 factors are used. As a supplementary experiment, the "stingy algorithm" has also been tried and exactly the same results have been obtained.

Subjective experiments have been conducted using the phoneme duration models with the top 1, 2, 3, 5, or 10 factors obtained by the quantification theory. Ten different sentence utterances have been synthesized for evaluation using the five duration models. Pair comparison tests between all the pairs of duration models have been conducted by 10 subjects. Each subject listened to three pairs of synthesized utterances selected randomly from the 10 sentences in each comparison condition and evaluated them in terms of naturalness. Figure 3 shows the preference scores. Although the best score is shown when using the top 5 factors, hypothesis testing based on the binomial distribution with the significance level of p-value ≤ 0.05 indicates that there is no significant difference among the four duration control methods with the top 2, 3, 5, or 10 factors.

4. SPEECH-RATE VARIABLE SYNTHESIS METHOD

This section proposes a method in which the phoneme duration can be controlled according to a speech rate given by a user. In order to implement this method, 1) three kinds of speech data with "fast", "normal", and "slow" speech rates are recorded, 2) a phoneme duration model is trained by the categorical multiple regression method for each speech rate, and 3) a duration model for arbitrary speech rate is made by interpolating the three models.

4.1. Database

Three speech databases at "fast", "normal", and "slow" speech rate, respectively, by a single male speaker "MTT" have been recorded. The normal speed database consists of 503 utterances reading the text used in the ATR continuous speech corpus. This database is also used for training triphone HMMs. The fast and slow databases respectively consist of 300 utterances reading a subset of the 503 sentences. The speaker was requested to read sentences as fast/slow as possible, keeping high intelligibility and naturalness. The fast and slow speed phoneme duration models have been trained using these 300 utterance database. Although the normal speed database has 503 utterances, the same 300 sentences have been used to make the normal speed duration model to maintain homogeneity.

In order to make a duration model at arbitrary speech rate, average mora length ML at speech rate s is defined as follows:

$$ML(s) = \frac{\text{Total duration except pause periods in the data}}{\text{Total number of mora in the data}} \quad (3)$$

The actual ML values, when s is $fast$, $normal$, and $slow$, are as follows:

$$\begin{aligned} ML(fast) &= 104.5 \text{ [ms]} \\ ML(normal) &= 149.6 \text{ [ms]} \\ ML(slow) &= 307.6 \text{ [ms]} \end{aligned}$$

4.2. Phoneme duration model generated by interpolation

The top five factors extracted by the Quantification Theory (Type 1), {`16`, `17`, `18`, `19`, `20`}, described in section 3.3, have been selected and three duration models with each speech rate have been trained. These five factors indicate the kind of five consecutive phonemes: two previous phonemes, the target phoneme, and two succeeding phonemes.

The duration model of the target speech rate st is obtained by

linearly interpolating the model parameter x_{fc} as follows:

$$x_{fc}(st) = \begin{cases} x_{fc}(normal)(1 - \frac{R(st)}{R(slow)}) \\ \quad + x_{fc}(slow)\frac{R(st)}{R(slow)} \quad (R(st) \geq 0) \\ x_{fc}(normal)(1 - \frac{R(st)}{R(fast)}) \\ \quad + x_{fc}(fast)\frac{R(st)}{R(fast)} \quad (R(st) < 0) \end{cases} \tag{4}$$

where $x_{fc}(s)$ indicates the score x_{fc} at speech rate s and $R(s)$ is a time-stretch ratio from the normal speech rate to the speech rate s. $R(s)$ is defined by the following equation:

$$R(s) = \frac{\log ML(s) - \log ML(normal)}{\log 2} \tag{5}$$

$R(s)$ values at the speech rates of the recorded databases are as follows:

$$\begin{array}{lcr} R(fast) & = & -0.52 \\ R(normal) & = & 0 \\ R(slow) & = & 1.04 \end{array}$$

4.3. Experiments

In order to evaluate the effectiveness of the interpolation, two target speech rates have been decided; "slightly fast" which is a middle speech rate between "fast" and "normal", and "slightly slow" which is a middle between "slow" and "normal". Therefore $R(slightly\ fast) = -0.26$ and $R(slightly\ slow) = 0.52$.

A rule-based method based on empirical knowledge has been made to compare with the proposed interpolation-based method. In the rule-based method, after computing each phoneme duration in a target sentence based on the duration model of the normal speech rate,

- the duration of vowels is multiplied by r,
- the duration of syllabic nasal(/N/) is multiplied by $\frac{r}{2}$, and
- the duration of other phonemes are maintained.

The coefficient r has been determined in each sentence so that both the proposed and rule-based methods produce a sentence of the same duration.

Subjective experiments were conducted using a pair comparison test between the proposed method and the rule-based method by 15 subjects. Utterances at four kinds of the speech rate: "fast", "slightly fast", "slightly slow", and "slow" were used. In the "fast" and "slow" speech rate conditions, the duration models were not interpolated. Each subject listened to pairs of three synthesized speech selected randomly from a set of 24 sentences at each speech rate and evaluated them in terms of naturalness. The text of the 24 synthesized speech were different from those used in training. Both methods used the same F_0 control model described in section 3.2 and the same pause length at every phrase boundary.

4.4. Experimental results

Figure 4 shows the preference scores of synthesized speech for the proposed method and the rule-based method at each speech rate. The proposed method shows better results than the rule-based method in all the experiments. The total preference score of our method is 65%. The hypothesis testing about this score indicates that our method is superior to the rule-based method with the significance level of p-value ≤ 0.05.

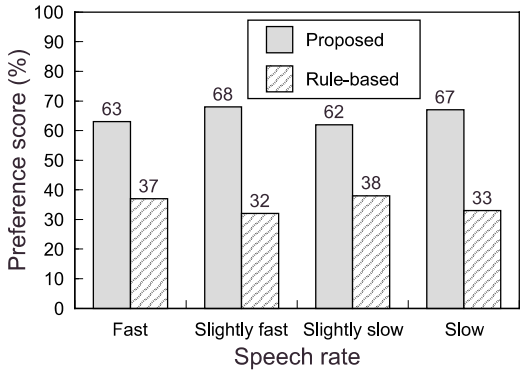

Fig. 4. Preference scores of synthesized speech by the proposed method to the rule-based method at various speech rate.

5. CONCLUSIONS

This paper has proposed a new speech synthesis method which can control phoneme duration according to arbitrary target speech rate. The phoneme duration control consists of two steps: 1) constructing three duration models at "fast", "normal", and "slow" speech rates, using Hayashi's Quantification Theory (Type 1) and 2) interpolating these models to create a model according to the target speech rate.

An analysis of the most effective factors for phoneme duration modeling indicates that the kinds of the estimation target phonemes and surrounding phonemes are especially important.

The proposed phoneme duration control method has been evaluated by subjective experiments at various speech rates using pair comparison tests between the proposed method and a rule-based method. The results show that the speech synthesized by the proposed method has significantly higher naturalness than the rule-based method.

Our future works include controlling the spectral parameters and the F_0 contour according to a target speech rate.

6. REFERENCES

[1] E. Moulines, et al., *Speech Commun.*, vol. 9, pp. 453–467 (1990-12).
[2] K. Tokuda, et al., *Proc. ICASSP 95*, vol. 1, pp. 660–663 (1995-5).
[3] T. Masuko, et al., *Proc. ICASSP 96*, vol. 1, pp. 389–392 (1996-5).
[4] C. Hayashi, *Ann. Inst. Statist. Math.* , vol. 3, no. 2, pp. 69–98 (1952).
[5] K. Hakoda, et al., *Proc. AVIOS 95*, pp. 65–72 (1995).
[6] H. Fujisaki, et al., *IEICE Trans. Fund.* , vol. E76-A, no. 11, pp. 1919–1926 (1993-11).
[7] T. Fukada, et al., *Proc. ICASSP 92*, vol. 1, pp. 137–140 (1992-3).
[8] W. Tachiwa, et al., *Proc. ASJ Spring Meeting 99*, vol. 1, pp. 239–240 (1999-3). (in Japanese)
[9] S. Imai, *Proc. ICASSP 83*, pp. 93–96 (1983-4).
[10] T. Sakayori, et al., *Proc. ASJ Autumn Meeting 86*, vol. 1, pp. 245–246 (1986-10). (in Japanese)
[11] M. Abe, et al., *Proc. ICASSP 92*, vol. 2, pp. 53–56 (1992-3).
[12] N. Kaiki, et al., *IEICE Trans. D-II*, vol. J83-D-II, no. 9, pp. 1853–1860 (2000-9). (in Japanese)
[13] N. Kaiki, et al., *Proc. ICSLP 90*, vol. 1, pp. 17–20 (1990-11).
[14] M. Yamada, et al., *Tech. report of IPSJ SLP*, vol. 2001, no. 100, pp. 15–20 (2001-10). (in Japanese)