DURATION MODELING FOR HMM-BASED SPEECH SYNTHESIS

Takayoshi Yoshimura†, Keiichi Tokuda†, Takashi Masuko††, Takao Kobayashi††and Tadashi Kitamura†

†Department of Computer Science
Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN
††Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Yokohama, 226-8502 JAPAN

ABSTRACT

This paper proposes a new approach to state duration modeling for HMM-based speech synthesis. A set of state durations of each phoneme HMM is modeled by a multi-dimensional Gaussian distribution, and duration models are clustered using a decision tree based context clustering technique. In the synthesis stage, state durations are determined by using the state duration models. In this paper, we take account of contextual factors such as stress-related factors and locational factors in addition to phone identity factors. Experimental results show that we can synthesize good quality speech with natural timing, and the speaking rate can be varied easily.

1. INTRODUCTION

For any text-to-speech synthesis system, controlling timing of the events in the speech signal is one of the difficult problems since there are many contextual factors (e.g., phone identity factors, stress-related factors, locational factors) that affect timing. Furthermore some factors affecting duration interact with one another. Recently, there have been proposed some approaches to controlling timing using statistical models such as linear regression [1], tree regression [2], MSR [3] which extends both linear and tree regressions, and sums-of-products model [4]. By using these techniques, rhythm and tempo of speech were successfully controlled with a small amount of free parameters.

On the other hand, we have proposed an HMM-based speech synthesis system in which the sequence of spectra is modeled by phoneme HMMs [5]. This synthesis system can synthesize speech with various voice characteristics by using a speaker adaptation technique [6], [7] or a speaker interpolation technique [8].

In this paper, we propose a new approach to controlling rhythm and tempo for the HMM-based speech synthesis system. In this approach, rhythm and tempo are controlled by state duration densities. State durations of each phoneme HMM is modeled by a multi-dimensional Gaussian distribution. Duration models are clustered using a decision tree based context clustering technique [10]. In the synthesis stage, state durations which maximize the state duration probability are determined from the state duration models and the total length of speech.

Since state durations are modeled by continuous distributions, our approach has the following advantages:

- The speaking rate of synthetic speech can be varied easily.
- There is no need for label boundaries when appropriate initial models are available since the state duration densities are estimated in the embedded training stage of phoneme HMMs.

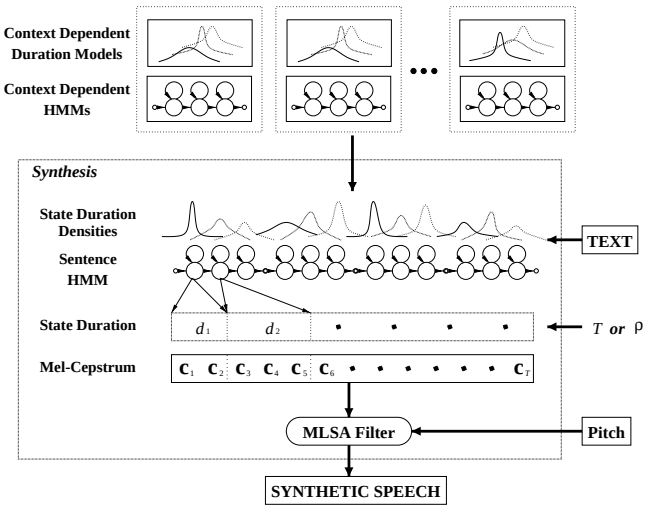

Figure 1: Speech synthesis system.

- Speaker individuality of synthetic speech can be varied by applying a speaker adaptation technique or a speaker interpolation technique to the HMMs and their state duration models.

In the following, we summarize the HMM-based speech synthesis system, and describe the technique for state duration modeling in Sections 2 and 3, respectively. Experimental results and discussions are also given in Section 4.

2. HMM-BASED SPEECH SYNTHESIS SYSTEM

The synthesis part of the HMM-based text-to-speech synthesis system is shown in Fig. 1.

HMMs and their duration models are context dependent models, where contextual factors which affect both spectra and state durations are taken into account.

In the training part, first, mel-cepstral coefficients are obtained from speech database using a mel-cepstral analysis technique [9], and delta coefficients are also calculated. Context dependent HMMs are trained using obtained coefficients. Using a decision-tree based context clustering technique [10], states of the context dependent HMMs are clustered, and the tied context dependent HMMs are

reestimated with the embedded training. Simultaneously, state durations are calculated on the trellis which is obtained in the embedded training stage, and modeled by Gaussian distributions. Finally, context dependent duration models are clustered by using the decision-tree based context clustering technique.

In the synthesis part, an arbitrarily given text to be synthesized is converted to a context-based label sequence. Then a sentence HMM is constructed by concatenating context dependent HMMs according to the label sequence. State durations of the sentence HMM are determined from the total length of speech T and the state duration densities. According to the obtained state durations, a sequence of mel-cepstral coefficients is generated from the sentence HMM by using a speech parameter generation algorithm [11], [12]. Finally, speech is synthesized from the generated mel-cepstral coefficients by the MLSA (Mel Log Spectrum Approximation) filter [9],[13].

3. STATE DURATION MODELING

In the HMM-based speech synthesis system described above, state duration densities were modeled by single Gaussian distributions estimated from histograms of state durations which were obtained by the Viterbi segmentation of training data. In this procedure, however, it is impossible to obtain variances of distributions for phonemes which appear only once in the training data.

In this paper, to overcome this problem, Gaussian distributions of state durations are calculated on the trellis which is made in the embedded training stage. State durations of each phoneme HMM are regarded as a multi-dimensional observation, and the set of state durations of each phoneme HMM is modeled by a multi-dimensional Gaussian distribution. Dimension of state duration densities is equal to number of state of HMMs, and nth dimension of state duration densities is corresponding to nth state of HMMs[1].

In the following sections, we describe training and clustering of state duration models, and determination of state duration in the synthesis part.

3.1. Training of State Duration Models

There have been proposed techniques for training HMMs and their state duration densities simultaneously, however, these techniques is inefficient because it requires huge storage and computational load. From this point of view, we adopt another technique for training state duration models.

State duration densities are estimated on the trellis which is obtained in the embedded training stage. The mean $\xi(i)$ and the variance $\sigma^2(i)$ of duration density of state i are determined by

$$\xi(i) = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)(t_1-t_0+1)}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)}, \quad (1)$$

$$\sigma^2(i) = \frac{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)(t_1-t_0+1)^2}{\sum_{t_0=1}^{T}\sum_{t_1=t_0}^{T}\chi_{t_0,t_1}(i)} - \xi^2(i), \quad (2)$$

[1]We assume the left-to-right model with no skip.

respectively, where $\chi_{t_0,t_1}(i)$ is the probability of occupying state i from time t_0 to t_1 and can be written as

$$\chi_{t_0,t_1}(i) = (1-\gamma_{t_0-1}(i)) \cdot \prod_{t=t_0}^{t_1}\gamma_t(i) \cdot (1-\gamma_{t_1+1}(i)), \quad (3)$$

where $\gamma_t(i)$ is the occupation probability of state i at time t, and we define $\gamma_{-1}(i) = \gamma_{T+1}(i) = 0$.

3.2. Decision-Tree Based Context Clustering

There are many combinations of contextual factors which affect duration such as phone identity factors, stress-related factors and locational factors. When we construct the state duration models taking account of many combinations of contextual factors, we expect to be able to obtain duration models which can predict natural timing accurately. However, as contextual factors increase, their combinations also increase exponentially. Therefore, model parameters with sufficient accuracy can not be estimated with limited training data. Furthermore, it is impossible to prepare speech database which includes all combinations of contextual factors; unseen contexts can not be prepared.

To overcome this problem, duration models are clustered using a decision-tree based context clustering technique. The decision tree is a binary tree, and in its each node, a question which splits contexts into two groups is prepared. All contexts can be found by traversing the tree, starting from the root node then selecting the next node depending upon the answer to a question about the current context. Therefore, if once the decision tree is constructed, unseen contexts can be prepared.

Our duration modeling technique using the decision tree is similar to the technique using CART [2]. Though the technique using CART can predict duration accurately, it can not control speaking rate easily because a discrete value is assigned to a leaf of the tree. In our approach, it is possible to control the speaking rate by assigning a multi-dimensional Gaussian distribution to a leaf of the tree.

3.3. Determination of State Duration

For a given speech length T, the goal is to obtain a state sequence $\mathbf{q} = \{q_1, q_2, \cdots, q_T\}$ which maximize

$$\log P(\mathbf{q}|\lambda, T) = \sum_{k=1}^{K}\log p_k(d_k) \quad (4)$$

under the constraint

$$T = \sum_{k=1}^{K} d_k, \quad (5)$$

where $p_k(d_k)$ is the probability of duration d_k in state k, and K is the number of states in HMM λ.

Since each duration density $p_k(d_k)$ is modeled by a single Gaussian distribution, state durations $\{d_k\}_{k=1}^{K}$ which maximize (4) are given by

$$d_k = \xi(k) + \rho\cdot\sigma^2(k) \quad (6)$$

$$\rho = \left(T - \sum_{k=1}^{K}\xi(k)\right) \Big/ \sum_{k=1}^{K}\sigma^2(k), \quad (7)$$

where $\xi(k)$ and $\sigma^2(k)$ are the mean and variance of the duration density of state k, respectively.

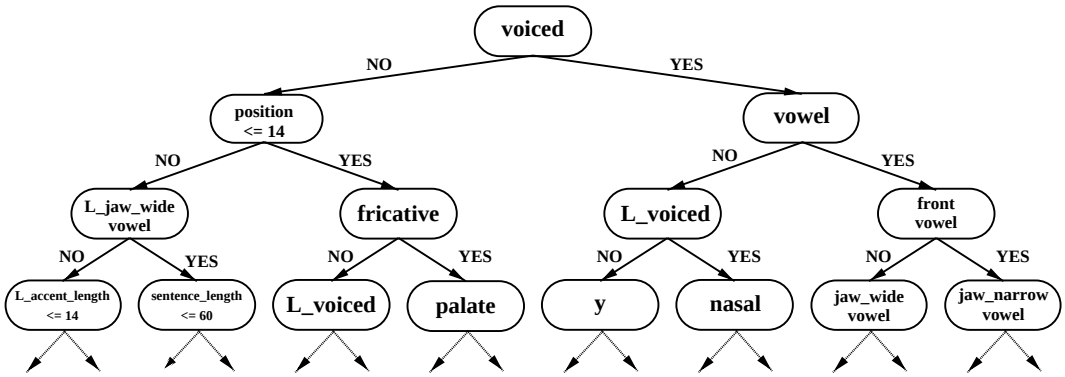

Figure 2: Decision tree for HMMs (1st state).

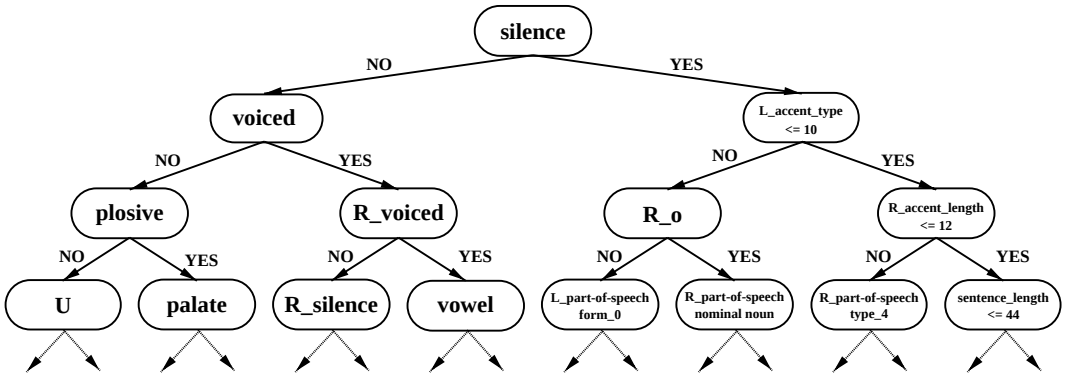

Figure 3: Decision tree for state duration models.

Since ρ is associated with T in (7), the speaking rate can be controlled by ρ instead of T. From (6), it can be seen that to synthesize speech with average speaking rate, ρ should be set to 0, that is, $T = \sum_{k=1}^{K} \xi(k)$, and the speaking rate becomes faster or slower when we set ρ to negative or positive value, respectively. It can also be seen that the variance $\sigma^2(k)$ represents "elasticity" of kth state duration.

4. EXPERIMENTS

We used phonetically balanced 450 sentences from ATR Japanese speech database for training. Speech signals were sampled at 16kHz and windowed by a 25ms Blackman window with a 5ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis. Feature vectors consisted of 25 mel-cepstral coefficients including the 0th coefficient, and their delta coefficients. We used 5-state left-to-right HMMs with single diagonal Gaussian output distributions.

Decision-tree based context clustering was applied to a set of context dependent HMMs. Then we estimated context dependent state duration models and applied context clustering to them. Following contextual factors which affect both spectra and state durations were taken into account:

- mora[2] count of sentence
- position of breath group in sentence
- mora count of {preceding, current, succeeding} breath group
- position of current accentual phrase in current breath group
- mora count and accent type of {preceding, current, succeeding} accentual phrase
- {preceding, current, succeeding} part of speech
- position of current phoneme in current accentual phrase
- {preceding, current, succeeding} phoneme

[2]A mora is a syllable-sized unit consisting of (consonant+) vowel.

The resultant set of HMMs and state duration models had 3,030 states and 2,984 distributions, respectively.

Examples of the decision tree for HMMs and their duration models are shown in Figs. 2 and 3, respectively. In these figures, "L_$*$" and "R_$*$" represent "preceding" and "succeeding", respectively. "silence" represents silence of head and tail of a sentence or pause. "L_accent_type $<= 10$" represents that the accent type of a preceding accentual phrase is from type zero to type ten. From these figures, with regard to spectra, it is seen that all models are much affected by phonetic identity. On the other hand, with regard to state duration, it can be seen that silence and pause models are much affected by accentual phrase and part-of-speech, and the other models are much affected by phonetic identity.

Fig. 4 shows generated spectra for a Japanese sentence which is not included in the training data, setting ρ to -0.1, 0, 0.1. Only the part corresponding to the first phrase "/t-o-k-a-i-d-e-w-a/", which means "in a city" in English, is shown in this figure. From the figure, it can be seen that some parts such as stationary parts of

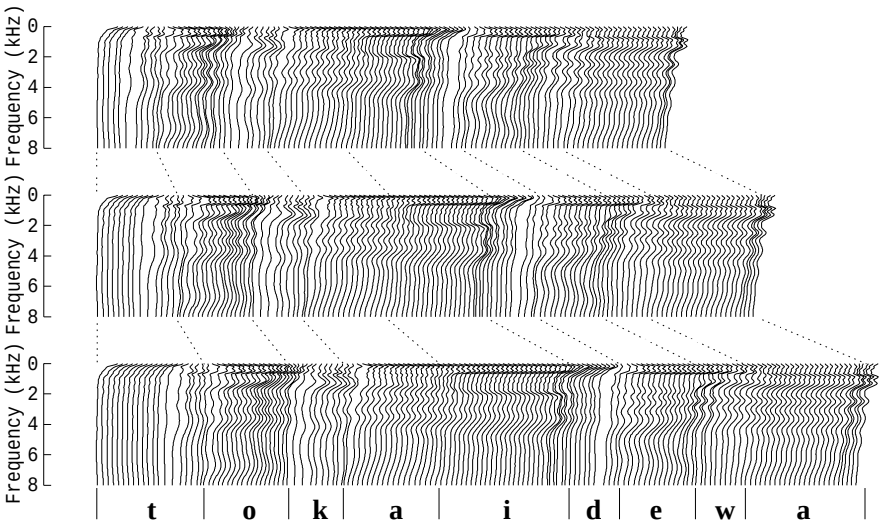

Figure 4: Generated spectra for an utterance "/t-o-k-a-i-d-e-w-a/" with different speaking rates (top : $\rho = -0.1$, middle : $\rho = 0$, bottom : $\rho = 0.1$).

vowels have elastic durations, and other parts such as explosives have fixed durations. From informal listening tests, we found that synthetic speech had a good quality with natural timing. Furthermore, we confirmed that synthetic speech could keep natural timing even if its speaking rate was changed in some degree.

5. CONCLUSION

In this paper, we described a state duration modeling technique for HMM-based speech synthesis, and constructed state duration models in which contextual factors that affect durations are taken into account. We synthesized speech using the constructed state duration models. As a results, we found that we can synthesize speech with natural timing and can control the speaking rate of the synthetic speech.

Future work will be directed towards investigation of contextual factors and conditions of the context clustering, and evaluation of synthetic speech. Building speech synthesis system which can deal with spectra, pitch [14], [15] and state duration in a unified framework, and synthesizing speech with various voice characteristics by applying speaker adaptation [6], [7] and speaker interpolation [8] techniques, are also our future works.

6. REFERENCES

1. N. Kaiki, K. Takeda, Y. Sagisaka: "Linguistic properties in the control of segmental duration for speech synthesis," *Talking Machines: Theories, Models, and Designs, Elsevier Science Publishers*, pp.255–263, 1992.
2. M. Riley: "Tree-based modelling of segmental duration," *Talking Machines: Theories, Models, and Designs, Elsevier Science Publishers*, pp.265–273, 1992.
3. N. Iwahashi and Y. Sagisaka: "Duration modelling with multiple split regression," *Proc. EUROSPEECH-93*, pp.329–332, 1993.
4. J. P. H. van Santen, C. Shih, B. Möbius, E. Tzoukermann and M. Tanenblatt: "Multi-lingual duration modeling," *Proc. EUROSPEECH-97*, vol5, pp.2651–2654, 1997.
5. T. Masuko, K. Tokuda, T. Kobayashi and S. Imai: "Speech synthesis from HMMs using dynamic features," *Proc. ICASSP-95*, pp.389–392, 1996.
6. T. Masuko, K. Tokuda, T. Kobayashi and S. Imai: "Voice characteristics conversion for HMM-based speech synthesis system," *Proc. ICASSP-97*, vol.3, pp.1611–1614, 1997.
7. M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi: "Speaker adaptation for HMM-based speech synthesis system using MLLR," *ESCA Workshop on Speech Synthesis*, 1998.
8. T. Yoshimura, K. Tokuda, T. Masuko T. Kobayashi and T. Kitamura: "Speaker interpolation in HMM-based speech synthesis system," *Proc. EUROSPEECH-97*, vol5, pp.2523–2526, 1997.
9. T. Fukada, K. Tokuda, T. Kobayashi and S. Imai: "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP-92*, vol.1, pp.137–140, 1992.
10. J. J. Odell: "The Use of Context in Large Vocabulary Speech Recognition," *PhD thesis, Cambridge University*, 1995.
11. K. Tokuda, T. Kobayashi and S. Imai: "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP-95*, pp.660–663, 1995.
12. K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and Satoshi Imai: "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features," *Proc. EUROSPEECH-95*, pp.757–760, 1995.
13. S. Imai: "Cepstral analysis synthesis on the mel frequency scale," *Proc. ICASSP-83*, pp.93–96, 1983.
14. N. Miyazaki, K. Tokuda, T. Masuko and T. Kobayashi: "An HMM based on multi-space probability distributions and its application to pitch pattern modeling," *IEICE Technical Report*, SP98-11, 1998, (in Japanese).
15. N. Miyazaki, K. Tokuda, T. Masuko and T. Kobayashi: "A study on pitch pattern generation using HMMs based on multi-space probability distributions," *IEICE Technical Report*, SP98-12, 1998, (in Japanese).