SIMULTANEOUS MODELING OF SPECTRUM, PITCH AND DURATION IN HMM-BASED SPEECH SYNTHESIS

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

†Nagoya Institute of Technology, Gokiso, Shouwa-ku, Nagoya, 466-8555 Japan
††Tokyo Institute of Technology, Nagatsuta, Midori-ku, Yokohama, 226-8502 Japan

E-mail: {yossie, tokuda, kitamura}@ics.nitech.ac.jp, {masuko, tkobayas}@ip.titech.ac.jp

ABSTRACT

In this paper, we describe an HMM-based speech synthesis system in which spectrum, pitch and state duration are modeled simultaneously in a unified framework of HMM. In the system, pitch and state duration are modeled by multi-space probability distribution HMMs and multi-dimensional Gaussian distributions, respectively. The distributions for spectral parameter, pitch parameter and the state duration are clustered independently by using a decision-tree based context clustering technique. Synthetic speech is generated by using an speech parameter generation algorithm from HMM and a mel-cepstrum based vocoding technique. Through informal listening tests, we have confirmed that the proposed system successfully synthesizes natural-sounding speech which resembles the speaker in the training database.

1. INTRODUCTION

Although most text-to-speech synthesis systems can synthesize speech with acceptable quality, they still cannot synthesize speech with various voice characteristics such as speaker individualities and emotions. To obtain various voice characteristics in text-to-speech synthesis systems based on the selection and concatenation of acoustical units, a large amount of speech data is necessary. However, it is difficult to collect, segment, and store it. From these points of view, in order to construct speech synthesis system which can generate various voice characteristics, we have proposed an HMM-based speech synthesis system [1].

Several HMM-based concatenative speech synthesis systems have also been proposed (e.g., [2], [3]). Our system, however, differs from them in that our system generates synthetic speech from HMMs themselves by using a speech parameter generation algorithm [4] and a mel-cepstrum based vocoding technique [5], [6]. In the parameter generation algorithm, by the inclusion of dynamic coefficients in the feature vector, the dynamic coefficients of the speech parameter sequence generated in synthesis are constrained to be realistic, as defined by the parameters of the HMMs. By transforming HMM parameters appropriately, voice characteristics of synthetic speech can be changed since the system generates speech from the HMMs. In fact, we have shown that we can change voice characteristics of synthetic speech by applying a speaker adaptation technique [7] or a speaker interpolation technique [8]. However, to change not only speaker individuality but also speaking style such as emotion expression, pitch and duration parameters have to be appropriately transformed.

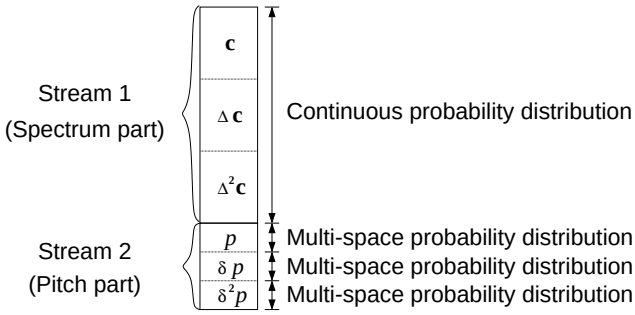


Figure 1. Feature vector.

In this paper, in order to apply speaker adaptation/ interpolation techniques to spectrum, pitch and state duration, simultaneously, and to synthesize speech with various voice characteristics, we construct a speech synthesis system in which spectrum, pitch and state duration are modeled simultaneously in a unified framework of HMM. In the system, pitch and state duration are modeled by multi-space probability distribution HMMs [9] and multi-dimensional Gaussian distributions [10], respectively. The feature vector of HMMs used in the system consists of two streams, i.e., the one for spectral parameter vector and the other for pitch parameter vector, and each phoneme HMM has its state duration densities. The distributions for spectral parameter, pitch parameter and state duration are clustered independently by using a decision-tree based context clustering technique [11]. In the context clustering, we take account of contextual factors which affect spectrum, pitch and duration such as phone identity factors, stress-related factors and locational factors.

The rest of this paper is structured as follows. Section 2 describes simultaneous modeling of spectrum, pitch and state duration. Section 3 describes the proposed text-to-speech synthesis system. Experimental results are presented in Section 4, and concluding remarks and our plans for future work are presented in the final section.

2. SIMULTANEOUS MODELING

2.1. Spectrum and Pitch Model

We use mel-cepstral coefficients as spectral parameter. Sequences of mel-cepstral coefficient vector, which are obtained from speech database using a mel-cepstral analysis technique [5], are modeled by continuous density HMMs. The mel-cepstral analysis technique enables speech to be re-synthesized from the mel-cepstral coefficients using the MLSA (Mel Log Spectrum Approximation) filter [6].

Pitch patterns are modeled by a hidden Markov model based on multi-space probability distribution (MSD-HMM) [9]. We cannot apply the conventional discrete or continuous HMMs to pitch pattern modeling since the observation sequence of pitch pattern is composed of one-dimensional continuous values and a discrete symbol which represents "unvoiced". The MSD-HMM includes discrete HMM and continuous mixture HMM as special cases, and further can model the sequence of observation vectors with variable dimension including zero-dimensional observations, i.e., discrete symbols. As a result, MSD-HMMs can model pitch patterns without heuristic assumption.

We construct spectrum and pitch models by using embedded training because the embedded training does not need label boundaries when appropriate initial models are available. However, if spectrum models and pitch models are embedded-trained separately, speech segmentations may be discrepant between them.

To avoid this problem, context dependent HMMs are trained with feature vector which consists of spectrum, pitch and their dynamic features (Fig. 1).

2.2. State Duration Model

State duration densities are modeled by single Gaussian distributions [10]. Dimension of state duration densities is equal to the number of state of HMM, and the n-th dimension of state duration densities is corresponding to the n-th state of HMMs[1]. 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.

There have been proposed techniques for training HMMs and their state duration densities simultaneously (e.g., [12]). However, these techniques require a large storage and computational load. In this paper, state duration densities are estimated by using state occupancy probabilities which are obtained in the last iteration of embedded re-estimation [10].

3. CONTEXT DEPENDENT MODEL

3.1. Contextual Factors

There are many contextual factors (e.g., phone identity factors, stress-related factors, locational factors) that affect spectrum, pitch and duration. In this paper, following contextual factors are 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

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

[2] A mora is a syllable-sized unit in Japanese.

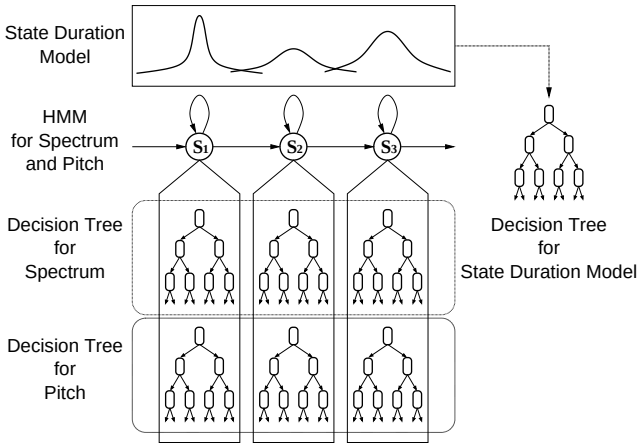


Figure 2. Decision trees.

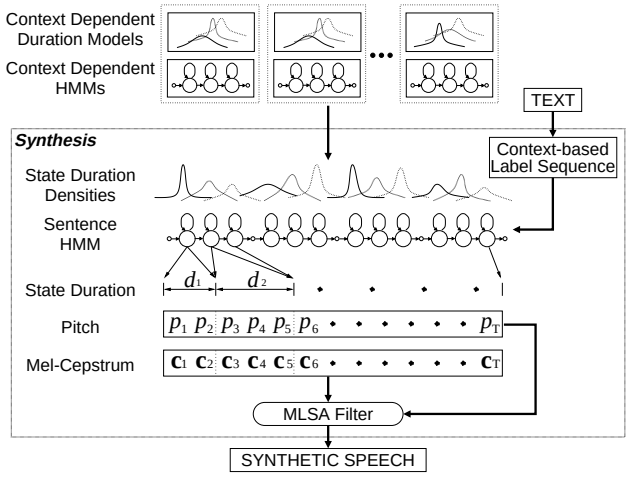


Figure 3. Synthesis part of the system.

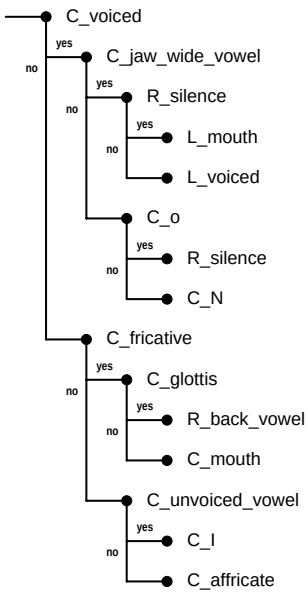
- {preceding, current, succeeding} part-of-speech
- position of current phoneme in current accentual phrase
- {preceding, current, succeeding} phoneme

Note that a context dependent HMM corresponds to a phoneme.

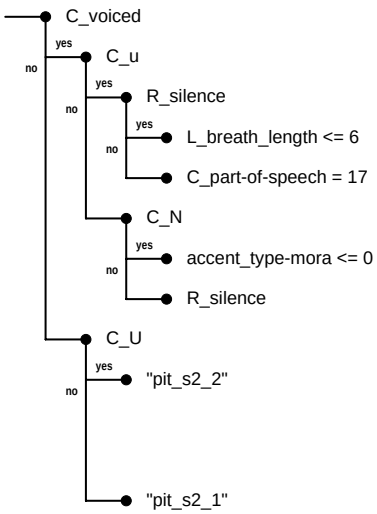
3.2. Decision-Tree Based Context Clustering

When we construct context dependent models taking account of many combinations of the above contextual factors, we expect to be able to obtain appropriate models. However, as contextual factors increase, their combinations also increase exponentially. Therefore, model parameters with sufficient accuracy cannot be estimated with limited training data. Furthermore, it is impossible to prepare speech database which includes all combinations of contextual factors.

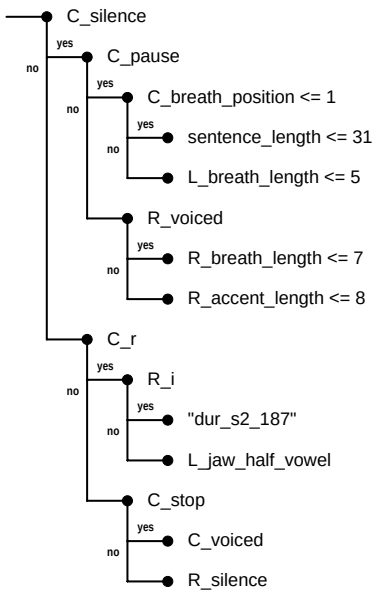
To overcome this problem, we apply a decision-tree based context clustering technique [11] to distributions for spectrum, pitch and state duration. The decision-tree based context clustering algorithm have been extended for MSD-HMMs in [13]. Since each of spectrum, pitch and duration have its own influential contextual factors, the distributions for spectral parameter and pitch parameter and the state duration are clustered independently (Fig. 2).



(a) Tree for Spectrum Model (1st state)



(b) Tree for Pitch Model (1st state)



(c) Tree for State Duration Model

Figure 4. Examples of decision trees.

4. TEXT-TO-SPEECH SYNTHESIS SYSTEM

The synthesis part of the HMM-based text-to-speech synthesis system is shown in Fig. 3. In the synthesis part, an arbitrarily given text to be synthesized is converted to a context-based label sequence. Then, according to the label sequence, a sentence HMM is constructed by concatenating context dependent HMMs. State durations of the sentence HMM are determined so as to maximize the likelihood of the state duration densities [10]. According to the obtained state durations, a sequence of mel-cepstral coefficients and pitch values including voiced/unvoiced decisions is generated from the sentence HMM by using the speech parameter generation algorithm [4]. Finally, speech is synthesized directly from the generated mel-cepstral coefficients and pitch values by the MLSA filter [5], [6].

5. EXPERIMENT

We used phonetically balanced 450 sentences from ATR Japanese speech database for training. Speech signals were sampled at 16 kHz and windowed by a 25-ms Blackman window with a 5-ms shift, and then mel-cepstral coefficients were obtained by the mel-cepstral analysis[3]. Feature vector consists of spectral and pitch parameter vectors. Spectral parameter vector consists of 25 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. Pitch parameter vector consists of log pitch, its delta and delta-delta. We used 3-state left-to-right HMMs with single diagonal Gaussian output distributions. Decision trees for spectrum, pitch and duration models were constructed as shown in Fig. 2. The resultant trees for spectrum models, pitch models and state duration models had 6,615, 1,877 and 201 leaves in total, respectively.

Fig. 4 shows examples of constructed decision trees for spectrum (a), pitch (b) and state duration (c). In these figures, "L_∗", "C_∗" and "R_∗" represent "preceding", "current" and "succeeding", respectively. "Silence" represents silence of head or tail of a sentence, or pause. Questions of breath group and accentual phrase are represented by "∗_breath_∗" and "∗_accent_∗", respectively. "Pit_s2_∗" and "dur_s2_∗" represent leaf nodes. From these figures, it is seen that spectrum models are much affected by phonetic identity, pitch models for "voiced" are much affected by accentual phrase and part-of-speech, and pitch models for "unvoiced" are clustered by a very simple tree. With regard to state duration models, 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. 5, 6 show generated spectra and pitch pattern, respectively, for a Japanese sentence "heikiNbairituwo sageta keisekiga aru" which is not included in the training data. Only the part corresponding to the first phrase "heikiNbairitu" is shown in Fig. 5.

The sound files[4] attached to this paper demonstrate our speech synthesis system. [Sound Y018S01.WAV] and [Sound Y018S02.WAV] are vocoded natural speech. [Sound Y018S11.WAV] and [Sound Y018S12.WAV] are speech generated from the system, and correspond to [Sound Y018S01.WAV] and [Sound Y018S02.WAV], respectively. It is observed that the system synthesizes natural-sounding speech which resembles the speaker in the training database.

Through informal listening tests, we have found that the stopping rule (a minimum frame occupancy at each leaf and a minimum gain in likelihood per splice) should be

[3]The source codes of the mel-cepstral analysis/synthesis can be found in *http://kt−lab.ics.nitech.ac.jp/˜tokuda/SPTK/* .

[4]These sound files were up-sampled to 22 kHz, and converted to WAV format. The latest sound files can be found in *http://kt−lab.ics.nitech.ac.jp/˜yossie/TTS/*.

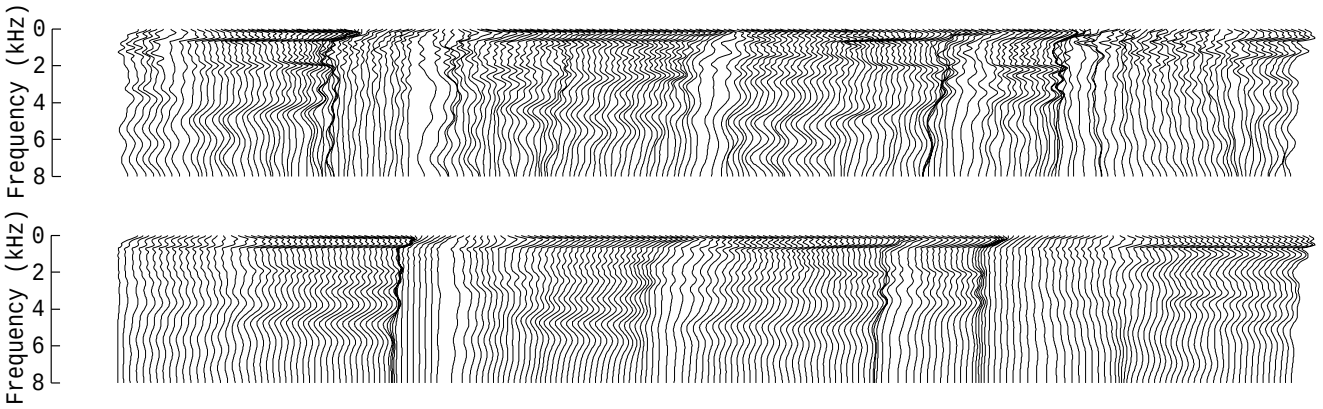


Figure 5. Generated spectra for a phrase "heikiNbairitsu"
(top: natural spectra, bottom: generated spectra).

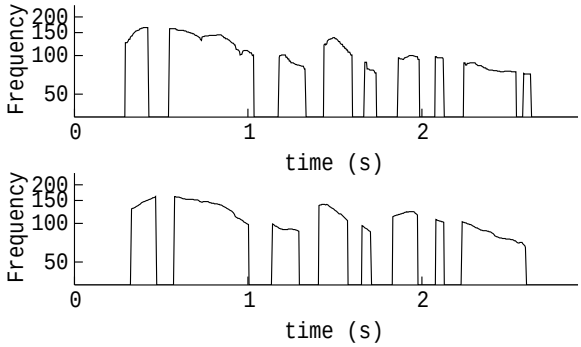


Figure 6. Generated pitch pattern for a sentence "heikiNbairitsuwo sageta keisekiga aru"
(top: natural pitch, bottom: generated pitch).

determined appropriately in decision tree construction. An overly large tree will be overspecialized to training data and generalize poorly. On the other hand, a overly small tree will model the data badly. Therefore we should introduce some stopping criterion or cross-validation method (e.g., [14]–[16]).

6. CONCLUSION

In this paper, we discribed an HMM-based speech synthesis system, in which spectrum, pitch and state duration are modeled simultaneously in a unified framework of HMM. As a result, it might be possible to synthesize speech with various voice characteristics, e.g., emotion expression, by applying speaker adaptation or speaker interpolation technique.

Future work will be directed towards investigation of contextual factors and conditions of the context clustering, and evaluation of synthetic speech. Synthesizing speech with various voice characteristics by applying speaker adaptation and speaker interpolation techniques is also our future work.

ACKNOWLEDGEMENT

This work was partially supported by the Ministry of Education, Science, Sports and Culture Japan, Grant-in-Aid for Scientific Research (B) 2, 1055125, 1998, Encouragement of Young Scientists, 0780226, 1998.

REFERENCES

[1] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. of ICASSP, pp.389–392, 1996.

[2] R. E. Donovan and E. M. Eide, "The IBM Trainable Speech Synthesis System," Proc. of ICSLP, vol.5, pp.1703–1706, 1998.

[3] M. Plumpe, A. Acero, H. Hon and X. Huang, "HMM-based Smoothing for Concatenative Speech Synthesis," Proc. of ICSLP, vol.6, pp.2751–2754, 1998.

[4] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S.Imai, "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features," Proc. of EUROSPEECH, pp.757–760, 1995.

[5] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. of ICASSP, vol.1, pp.137–140, 1992.

[6] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," Proc. of ICASSP, pp.93–96, 1983.

[7] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR," Proc. of The Third ESCA/COCOSDA workshop on Speech Synthesis, pp.273–276, 1998.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Speaker Interpolation in HMM-Based Speech Synthesis System," Proc. of EUROSPEECH, vol.5, pp.2523–2526, 1997.

[9] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," Proc. of ICASSP, 1999.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System," Proc. of ICSLP, vol.2, pp.29–32, 1998.

[11] J. J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, Cambridge University, 1995.

[12] S. E. Levinson, "Continuously Variable Duration Hidden Markov Models for Speech Analysis," Proc. of ICASSP, pp.1241–1244, 1986.

[13] N. Miyazaki, K. Tokuda, T. Masuko and T. Kobayashi, "A Study on Pitch Pattern Generation using HMMs Based on Multi-space Probability Distributions," Technical Report of IEICE, SP98-12, 1998 (in Japanese).

[14] H. J. Nock, M. J. F. Gales and S. J. Young, "A Comparative Study of Methods for Phonetic Decision-Tree State Clustering," Proc. of EUROSPEECH, pp.111–115, 1997.

[15] K. Shinoda and T. Watanabe, "Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle," Proc. of ICASSP, pp.717–720, 1996.

[16] W. Chou and W. Reichl, "Decision Tree State Tying Based on Penalized Bayesain Information Criterion," Proc. of ICASSP, pp.345–348, 1999.