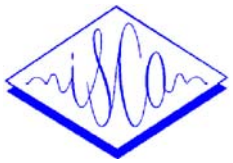


Speech Prosody 2004
Nara, Japan
March 23-26, 2004

# HMM-Based Speech Synthesis with Various Speaking Styles Using Model Interpolation

*Makoto Tachibana, Junichi Yamagishi, Koji Onishi, Takashi Masuko & Takao Kobayashi*

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
{Makoto.Tachibana; Junichi.Yamagishi; masuko; takao.kobayashi}@ip.titech.ac.jp

## Abstract

This paper presents an approach to realizing various speaking styles and emotional expressions using a model interpolation technique in HMM-based speech synthesis. In the approach, we synthesize speech with an intermediate speaking style between representative speaking styles from a model obtained by interpolating representative style models. We chose three styles, "reading," "joyful," and "sad," as representative styles, and synthesized speech from models obtained by interpolating two models for every combination of two styles. From a result of a subjective similarity evaluation, it is shown that speech generated from an interpolated model has a speaking style in between two representative speaking styles.

## 1. Introduction

In text-to-speech synthesis, to change speaking style and emotional expression of the synthetic speech arbitrarily with maintaining its naturalness, it is required that prosodic features as well as spectral features are controlled properly. Moreover, since prosodic features are more or less related to spectral features, it is undesirable to control these features independently. This would make the problem difficult for generating natural sounding speech in accordance with the specified speaking style and/or emotional expression. In fact, although many researchers have attempted to add emotional expression to synthetic speech (for example, see [1] and references therein), most of the approaches are rule-based and do not always take account of the relationship between spectrum and prosody.

In this paper, we describe an approach to resolving this problem. The approach is based on a speaking style and emotional expression modeling technique for HMM-based speech synthesis [2]. In this technique, speaking styles and emotional expressions are fully statistically modeled, and synthetic speech is generated without using rules controlling prosody and other parameters. Since spectral and prosodic features are modeled simultaneously in each speaking style model, we can incorporate the relationship between spectrum and prosody into the speaking style control process implicitly. Moreover it has been shown that we can synthesize speech with similar speaking styles and emotional expressions to those of the recorded speech.

Here we investigate a method for synthesizing speech with an intermediate speaking style between two different speaking styles by applying a model interpolation technique [3]. We will refer to one of speaking styles or emotional expressions as the "style." In this paper, we choose three styles, namely "reading," "joyful," and "sad" styles, and synthesize speech from models obtained by interpolating two models for every combination of two styles. From a result of a subjective experiment we show that speech generated from an interpolated model has a speaking style in between two speaking styles.

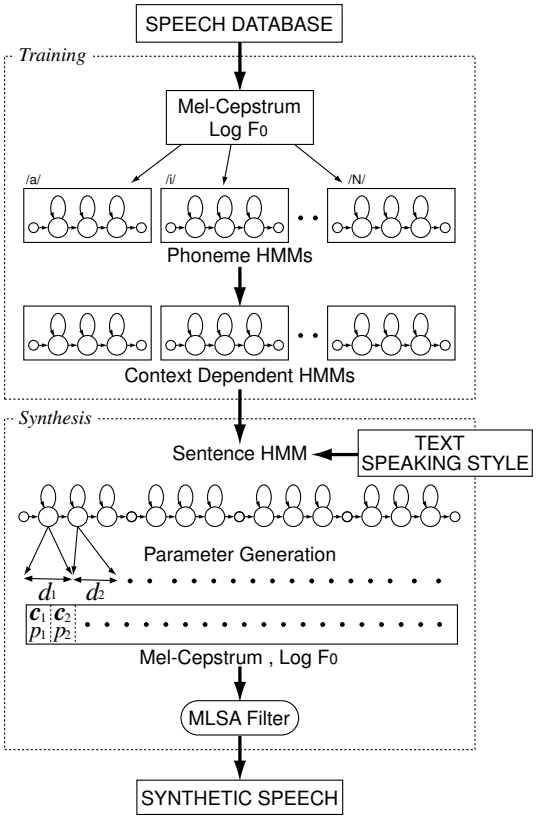



Figure 1: *A block diagram of an HMM-based speech synthesis system.*

## 2. HMM-Based Speech Synthesis with Various Speaking Styles

### 2.1. Overview of HMM-Based Speech Synthesis

A block diagram of the HMM-based TTS system [4] is shown in Fig. 1. The system consists of two stages, the training stage and the synthesis stage.

In the training stage, phoneme HMMs are trained using speech database. Spectrum and $F_0$ are modeled by multi-stream HMMs in which output distributions for spectral and $F_0$

parts are modeled using continuous probability distribution and multi-space probability distribution (MSD) [5], respectively. To model variations of spectrum and $F_0$, phonetic, prosodic, and linguistic contextual factors, such as phoneme identity factors, stress related factors, and locational factors, are taken into account. Then, a decision tree based context clustering technique [6, 7] is separately applied to the spectral and $F_0$ parts of the context dependent phoneme HMMs. Finally, state durations are modeled by multi-dimensional Gaussian distributions, and the state clustering technique is applied to the duration models.

In the synthesis stage, first, an arbitrarily given text is transformed into a context dependent phoneme label sequence. According to the label sequence, a sentence HMM is constructed by concatenating context dependent phoneme HMMs. Phoneme durations are determined using state duration distributions, and then spectral and $F_0$ parameter sequences are obtained based on ML criterion from the sentence HMM. Finally, by using the MLSA filter [8], speech is synthesized from the generated mel-cepstral and $F_0$ parameter sequences.

### 2.2. Speaking Style Modeling in HMM-Based Speech Synthesis

In [2], we have proposed two methods for modeling speaking styles called "style dependent modeling" and "style mixed modeling." In style dependent modeling, each style is individually modeled. On the other hand, in style mixed modeling, speaking styles and emotional expressions are treated as a contextual factor as well as phonetic and linguistic factors, and all styles are modeled by a single acoustic model simultaneously. We have shown that both modeling methods have almost the same performance, and that it is possible to synthesize speech which have similar styles to those of recoded speech. Although style mixed modeling has been shown to be able to reduce the total number of output distributions, we adopt style dependent modeling in this paper because it is easy to add or remove styles without retraining the whole model.

## 3. Interpolation of Speaking Style Models

It has been shown in [3] that it is possible to synthesize speech with intermediate voice characteristics between two speakers by interpolating two speakers' models. In this paper, we apply the model interpolation technique to speaking style modeling to synthesize speech with an intermediate speaking style between representative speaking styles.

In [3], three interpolation methods are proposed; (a) interpolation between observations, (b) interpolation between output distributions of HMM states taking account of state occupancies, and (c) interpolation based on Kullback information measure. Among these methods, we adopted the simplest method, i.e., method (a), in the following.

Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be models of $N$ representative speaking styles $S_1, S_2, \ldots, S_N$, and $\tilde{\lambda}$ be a model of a speaking style $\tilde{S}$ obtained by interpolating $N$ representative style models. When an observation vector $\tilde{\boldsymbol{o}}$ of the speaking style $\tilde{S}$ is obtained by linearly interpolating observation vectors $\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_N$ of the representative speaking styles as follows,

$$\tilde{\boldsymbol{o}} = \sum_{k=1}^{N} a_k \boldsymbol{o}_k, \tag{1}$$

where $\sum_{k=1}^{N} a_k = 1$, a mean vector $\tilde{\boldsymbol{\mu}}$ and a covariance matrix $\tilde{\boldsymbol{U}}$ of a Gaussian output pdf $p(\tilde{\boldsymbol{o}}) = \mathcal{N}(\tilde{\boldsymbol{o}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{U}})$ is calculated by

$$\tilde{\boldsymbol{\mu}} = \sum_{k=1}^{N} a_k \boldsymbol{\mu}_k, \tag{2}$$

$$\tilde{\boldsymbol{U}} = \sum_{k=1}^{N} {a_k}^2 \boldsymbol{U}_k, \tag{3}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{U}_k$ are a mean vector and a covariance matrix of an output pdf of speaking style $S_k$, respectively.

If the models $\lambda_k$ $(1 \leq k \leq N)$ of representative speaking styles have a tying structure common to all models, it is possible to obtain the interpolated model $\tilde{\lambda}$ by interpolating $\lambda_k$ directly. In general, however, the models $\lambda_k$ have different structure from each other when the context clustering is independently performed for each speaking style model in the training stage. Consequently, it is difficult to obtain $\tilde{\lambda}$ by interpolating $\lambda_k$ taking account of model structure. To avoid this problem, in the synthesis stage, we first generate $N$ pdf sequences from $\lambda_k$ independently, and then obtain a pdf sequence corresponding to $\tilde{\lambda}$ by interpolating these $N$ pdf sequences. Finally, a speech parameter sequence is generated from the interpolated pdf sequence.

## 4. Experiments

### 4.1. Speech Database and Experimental Conditions

Although there exist a wide variety of speaking styles and emotions in real speech, it is not easy to collect them completely. As the first step toward modeling and synthesis of expressive speech, we chose three speaking styles, namely "reading," "joyful," and "sad" as the representative styles. We used the speech database [2] which contains a set of phonetically balanced 503 sentences of ATR Japanese speech database uttered by a male speaker MMI and a female speaker FTY for each speaking style.

We used 42 phonemes including silence and pause, and the following contextual factors were taken into account:

- the number of morae in sentence
- position of breath group in sentence
- the number of morae in {preceding, current, succeeding} breath group
- position of current accentual phrase in current breath group
- the number of morae and accent type in {preceding, current, succeeding} accentual phrase
- {preceding, current, succeeding} part-of-speech
- position of current mora in current accentual phrase
- difference between position of current mora and accent type
- {preceding, current, succeeding} phoneme

It is noted that these contextual factors are the same as [4] in which only reading style is taken into account.

Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis [8, 9]. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. We used 5-state left-to-right HMMs. The representative style models were trained using 450 sentences for each style.

Table 1: *Classification result of styles of synthesized speech using representative models.*

| Synthetic | Classification Rate (%) | | | |
|---|---|---|---|---|
| Speech | Read. | Joyful | Sad | Other |
| Read. | 98.3 | 0.0 | 0.0 | 1.7 |
| Joyful | 1.1 | 94.9 | 0.0 | 4.0 |
| Sad | 0.6 | 0.0 | 94.9 | 4.5 |

We obtained interpolated style speech samples for three combinations, that is, "reading" and "joyful," "reading" and "sad," and "joyful" and "sad." Here we denote the new speaking style interpolated between styles A and B as (A, B). Interpolation ratio is set to $1:1$ ($(a_A, a_B) = (0.5, 0.5)$) for all cases.

### 4.2. Subjective Evaluation Results

We first conducted a classification test for styles of generated speech from the representative models. Subjects were eleven males, and asked which style, namely "reading," "joyful," and "sad", the test speech sounded. It it noted that test speech was classified into "Other" when it was thought to be classified into none of the above tree styles. For each subject, eight test sentences were chosen at random from 53 test sentences which were not contained in the training data. Table 1 shows the classification rates for synthesized speech. It can be seen from the result that it is possible to synthesize speech with similar styles to those of the recorded speech.

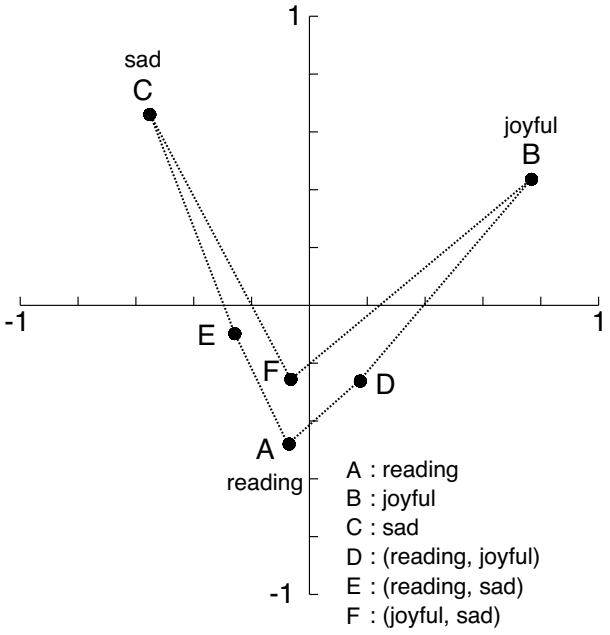
We then conducted a subjective test to evaluate similarity of speech samples between the following representative and interpolated speaking styles:

- A: reading
- B: joyful
- C: sad
- D: (reading, joyful)
- E: (reading, sad)
- F: (joyful, sad)

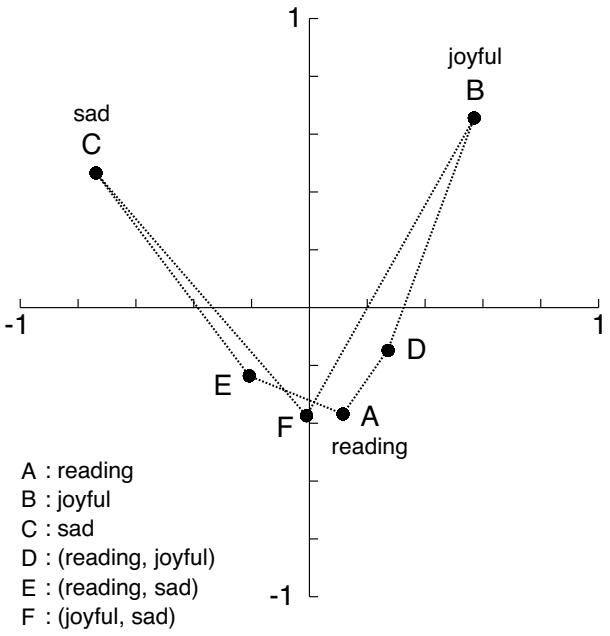
Subjects were presented two speech samples chosen from the above six styles in random order, and asked to evaluate similarity of each pair in a five point scale in which "5" means very similar and "1" means quite different. For each subject, four test sentences were chosen at random from 53 test sentences which were not included in the training data. Subjects were eight males.

From the result of similarity evaluation, we placed six styles in a 2-dimensional space according to the similarities between styles by using the Hayashi's quantification theory type IV [10]. Figure 2 shows the relative similarity distance between speaking styles. Figure 2 (a) shows the result for male speaker MMI, and (b) shows the result for female speaker FTY. In this figure, it is thought that the horizontal axis corresponds to the degree of pleasure, and the vertical axis corresponds to intensity of emotional expression.

From this figure, it can be seen that interpolated speaking styles between "reading" and "joyful" or "sad" (D and E) are placed in between representative speaking styles. From this result, it is thought that we can synthesize speech with a speaking style in between two representative speaking styles by using model interpolation technique. It can also be thought that




(a)male speaker MMI




(b)female speaker FTY

Figure 2: *Evaluation of similarity between speaking styles.*

since interpolated speaking style between "joyful" and "sad" (F) is placed near the "reading" style (A), "joyful" and "sad" styles have opposite features putting "reading" style between these two styles in the model parameter space.

## 5. Conclusions

In this paper, we investigated a technique for synthesizing speech with a speaking style in between two different speaking styles by applying a model interpolation technique. From the result of the subjective experiment, it was shown that speech

generated from an interpolated model has a speaking style in between two representative speaking styles.

Future work will focus on investigation using other speakers and speaking styles. Investigation of other approach to synthesizing speech with various voice characteristics and speaking styles, such as an eigenvoice technique [11], is also our future work.

## 6. References


[1] Schröder, M., 2001. Emotional speech synthesis: a review. *Proc. EUROSPEECH-2001*, 1, 561-564.

[2] Yamagishi, J.; Onishi, K.; Masuko, T.; Kobayashi, T., 2003. Modeling of various speaking styles and emotions for HMM-based speech synthesis. *Proc. EUROSPEECH-2003*, 3, 2461-2464.

[3] Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T., 2000. Speaker interpolation for HMM-based speech synthesis system. *J. Acoust. Soc. Jap. (E)*, 21(4), 199-206.

[4] Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *Proc. EUROSPEECH-99*, 5, 2347-2350.

[5] Tokuda, K.; Masuko, T.; Miyazaki, N.; Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. *Proc. ICASSP-99*, 1, 229-232.

[6] Young, S.J.; Odell, J.; Woodland P., 1994. Tree-based state tying for high accuracy acoustic modeling. *Proc. ARPA Human Language Technology Workshop*, 307-312.

[7] Shinoda, K.; Watanabe, T., 2000. MDL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Jpn. (E)*, 21(2), 79-86.

[8] Fukada, T.; Tokuda, K.; Kobayashi, T.; Imai S., 1992. An adaptive algorithm for mel-cepstral analysis of speech. *Proc. ICASSP-92*, 1, 137-140.

[9] Tokuda, K.; Kobayashi, T.; Fukada, T.; Saito, H.; Imai, S., 1991. Spectral estimation of speech based on mel-cepstral representation. *IEICE Trans. Fundamentals (Japanese Edition)*, J74-A(8), 1240-1248.

[10] Hayashi, C., 1952. On the prediction of phenomena from mathematicostatistical point of view. *Annals of the Institute of Statistical Mathematics*, 3, 69-98.

[11] Shichiri, K.; Sawabe, A.; Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T., 2002. Eigenvoices for HMM-based speech synthesis. *Proc. ICSLP-2002*, 2, 1269-1272.