

# SPEAKER INTERPOLATION IN HMM-BASED SPEECH SYNTHESIS SYSTEM

Takayoshi Yoshimura<sup>1</sup>, Takashi Masuko<sup>2</sup>, Keiichi Tokuda<sup>1</sup>, Takao Kobayashi<sup>2</sup> and Tadashi Kitamura<sup>1</sup>

<sup>1</sup>Department of Computer Science, Nagoya Institute of Technology, Nagoya 466, Japan

<sup>2</sup>Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama 226, Japan

E-mail: yossie@ics.nitech.ac.jp, masuko@pi.titech.ac.jp, tokuda@ics.nitech.ac.jp,

tkobayas@pi.titech.ac.jp, kitamura@ics.nitech.ac.jp

## ABSTRACT

This paper describes an approach to voice characteristics conversion for HMM-based text-to-speech synthesis system by using speaker interpolation. An HMM interpolation technique is derived from a probabilistic distance measure for HMMs, and used to synthesize speech with untrained speaker's characteristics by interpolating HMM parameters among some representative speakers' HMM sets. The result of subjective experiments shows that the characteristics of synthesized speech is gradually changed from one's to the other's by changing the interpolation ratio.

## 1. INTRODUCTION

Although most text-to-speech synthesis systems can synthesize speech with acceptable quality, it 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.

For the purpose of synthesizing speech with various voice characteristics, we have proposed an algorithm for speech parameter generation from HMMs [1], [2], in which speech parameter (e.g., mel-cepstral coefficients) sequence is determined so that the output probability of the speech parameter sequence is maximized for given HMM, and applied the algorithm to an HMM-based speech synthesis system [3]. It was shown in [4], [5] that the HMM-based speech synthesis system can synthesize speech with target speaker's voice characteristics applying a speaker adaptation technique used in speech recognition.

This paper proposes a speaker interpolation technique for the HMM-based speech synthesis system to synthesize speech with untrained speaker's characteristics by interpolating HMM parameters among some representative speakers' HMM sets. By changing the interpolation ratio, we can gradually change the characteristics of synthesized speech from one's to the other's. Although the idea is similar to that of [6], in the proposed method we can use mathematically-well-defined statistical distance measures since each speech unit is represented by an HMM. Listening tests show

that the proposed algorithm successfully generates new speakers' voice characteristics in the case where two representative HMMs are trained by a male and a female speakers' speech data; the characteristics of synthesized speech is in between the male and female speakers, and can be gradually changed from one's to the other's according to interpolation ratio.

## 2. SPEECH SYNTHESIS SYSTEM

The proposed text-to-speech synthesis system is based on the speech parameter generation algorithm from HMMs [1], [2], and a mel-cepstral speech analysis/synthesis technique [7], [8]. A block diagram of the proposed text-to-speech synthesis system is shown in Fig.1, which is almost equivalent to the previously proposed system [3] except that multiple speaker's HMM sets are trained and interpolated to generate a new speaker's HMM set. The procedure can be summarized as follows:

1. Training representative HMM sets
  - (a) Select several representative speakers  $S_1, S_2, \dots, S_N$  appropriately from speech database, and repeat (b), (c) for each speaker.
  - (b) Obtain mel-cepstral coefficients from speech of the representative speaker by mel-cepstral analysis [8].
  - (c) Train phoneme HMM set  $\lambda_k$  using mel-cepstral coefficients, and their deltas and delta-deltas.
2. Interpolating representative HMM sets
  - (a) Generate a new phoneme HMM set  $\lambda$  by interpolating between the representative speakers' phoneme HMM sets  $\lambda_1, \lambda_2, \dots, \lambda_N$  with an arbitrary interpolation ratio  $a_1, a_2, \dots, a_N$  based on the method described in the next section.
3. Speech Synthesis from Interpolated HMM
  - (a) Transform the text to be synthesized into a phoneme sequence, and concatenate the interpolated phoneme HMMs according to the phoneme sequence.
  - (b) Generate speech parameter sequence from the sentence HMM by using speech parameter generation algorithm [1], [2], [3].
  - (c) Synthesize speech from the generated mel-cepstral coefficients by using the MLSA (Mel Log Spectral Approximation) filter [7].

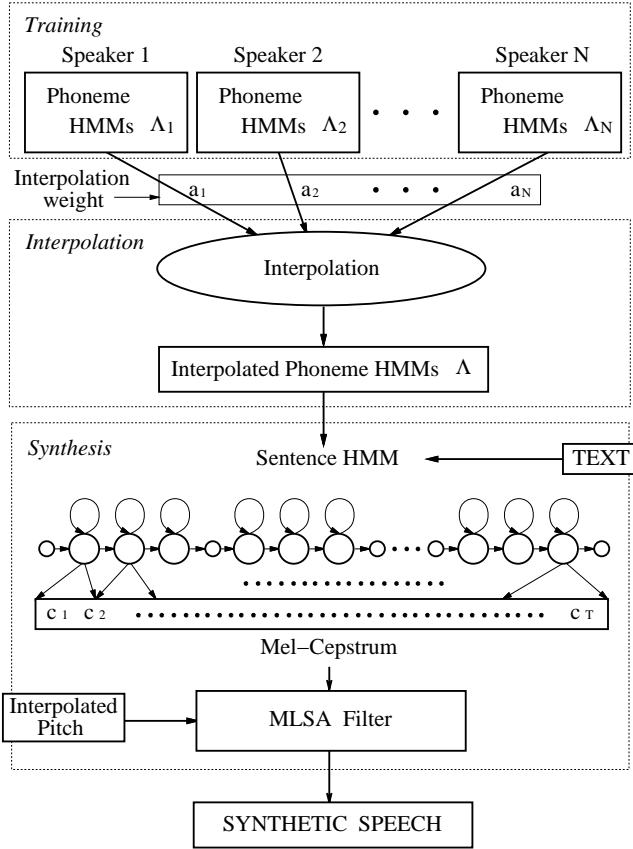


Figure 1. Block diagram of speech synthesis system.

In mel-cepstral analysis [8], the mel-cepstral coefficients, i.e., frequency-transformed cepstral coefficients are determined by maximizing  $P(\mathbf{x}|\mathbf{c})$ , where  $\mathbf{c}$  is the mel-cepstral coefficients, and  $\mathbf{x}$  is the input speech sequence assumed to be Gaussian. The transfer function  $D(z)$  defined by the mel-cepstral coefficients is approximated by the MLSA filter [7] with sufficient accuracy. By exciting the MLSA filter by pulse train or white noise according to a pitch contour, we can synthesize speech directly from the mel-cepstral coefficients with a small computational cost.

### 3. INTERPOLATION

Fig. 2 shows a space of speaker individuality. Representative speakers  $S_1, S_2, \dots, S_N$  are represented by HMMs,  $\lambda_1, \lambda_2, \dots, \lambda_N$ , respectively. When a speaker  $S$  is represented by an HMM  $\lambda$ , the distance between the interpolated speaker  $S$  and each representative speaker  $S_k$  is measured by Kullback information measure between  $\lambda$  and  $\lambda_k$ :

$$I(\lambda, \lambda_k) = E_{\mathbf{O}} \left[ P(\mathbf{O}|\lambda) \log \frac{P(\mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda_k)} \right]. \quad (1)$$

Thus, when we consider interpolating between  $N$  HMMs  $\lambda_1, \lambda_2, \dots, \lambda_N$  with weights  $a_1, a_2, \dots, a_N$ , it is reasonable to determine interpolated HMM  $\lambda$  in such

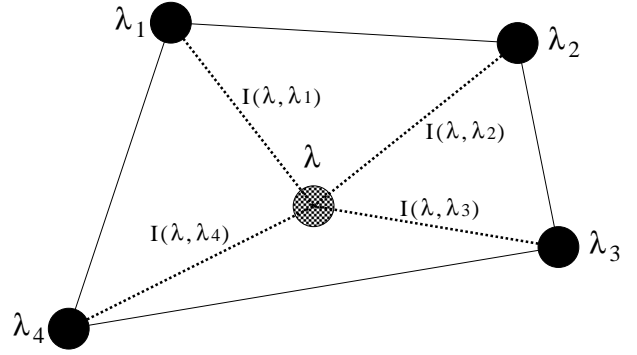


Figure 2. A space of speaker individuality represented by HMMs.

a way that  $\lambda$  maximizes a cost function

$$\varepsilon = \sum_{k=1}^N a_k I(\lambda, \lambda_k). \quad (2)$$

In this paper, we assume that HMMs to be interpolated have the same topology<sup>1</sup>. Under this assumption, interpolation between HMMs is equivalent to interpolation between corresponding states when state-transition probabilities are ignored. If we assume that each HMM state has a single Gaussian output probability density, the problem is reduced to interpolation between  $N$  Gaussian pdfs,  $p_k(\mathbf{o}) = \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_k, \mathbf{U}_k)$ ,  $k = 1, 2, \dots, N$ , where  $\boldsymbol{\mu}_k$  and  $\mathbf{U}_k$  denote mean vector and covariance matrix, respectively. Consequently, the interpolated pdf  $p(\mathbf{o}) = \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \mathbf{U})$  is determined by minimizing

$$\varepsilon = \sum_{k=1}^N a_k I(p, p_k) \quad (3)$$

with respect to  $\boldsymbol{\mu}$  and  $\mathbf{U}$ , where the Kullback information measures can be written as

$$\begin{aligned} I(p, p_k) &= E_{\mathbf{o}} \left[ \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \mathbf{U}) \log \frac{\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \mathbf{U})}{\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_k, \mathbf{U}_k)} \right] \\ &= \frac{1}{2} \left\{ \log \frac{|\mathbf{U}_k|}{|\mathbf{U}|} + \right. \\ &\quad \left. \text{tr} [\mathbf{U}_k^{-1} \{ (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})' + \mathbf{U} \}] + \mathbf{I} \right\}. \quad (4) \end{aligned}$$

As a result,  $\boldsymbol{\mu}$  and  $\mathbf{U}$  are determined by

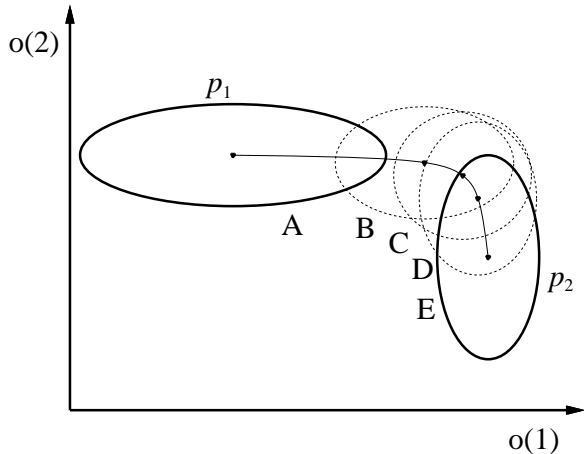
$$\boldsymbol{\mu} = \left( \sum_{k=1}^N a_k \mathbf{U}_k^{-1} \right)^{-1} \sum_{k=1}^N a_k \mathbf{U}_k^{-1} \boldsymbol{\mu}_k \quad (5)$$

$$\mathbf{U} = \left( \sum_{k=1}^N a_k \mathbf{U}_k^{-1} \right)^{-1} \sum_{k=1}^N a_k, \quad (6)$$

respectively.

Fig. 3 shows Gaussian distributions generated by interpolation between Gaussian distributions  $p_1$  and

<sup>1</sup>Distributions could be tied.



**Figure 3. Interpolation between two Gaussian distributions  $p_1$  and  $p_2$  with interpolation ratios** A :  $(a_1, a_2) = (1, 0)$ , B :  $(a_1, a_2) = (0.75, 0.25)$ , C :  $(a_1, a_2) = (0.5, 0.5)$ , D :  $(a_1, a_2) = (0.25, 0.75)$ , E :  $(a_1, a_2) = (0, 1)$ .

$p_2$  with two dimensional diagonal covariances. In this figure, each ellipse represents the contour corresponding to the standard deviation (squared variance) of a Gaussian distribution, and each dot represents the mean vector of the distribution. From the figure, it is seen that two distributions are interpolated appropriately in the sense that the interpolated distribution  $p$  reflects the statistical information of  $p_1$  and  $p_2$ , i.e., covariances of those.

#### 4. EXPERIMENTS

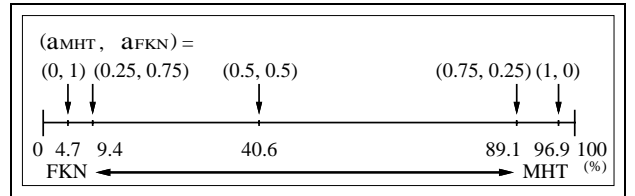
By analyzing the result of ABX listening tests and subjective experiments of similarity using Hayashi's fourth method of quantification[9], it is investigated whether the characteristics of synthesized speech from the interpolated HMM set is in between two trained speakers.

We used phonetically balanced 503 sentences from ATR Japanese speech database for training. Speech signals were windowed by a 25.6ms Hamming window with a 5ms shift, and then mel-cepstral coefficients were obtained by mel-cepstral analysis. The feature vectors consisted of 13 mel-cepstral coefficients including the 0th coefficient, and their delta and delta-delta coefficients.

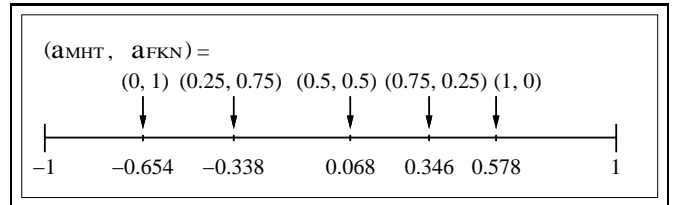
We used 5-state left-to-right triphone models with single Gaussian diagonal output distribution. Decision-tree based model clustering was applied to each set of triphone models, and the resultant set of tied triphone models has approximately 2,800 distributions.

We trained 2 HMM sets using 503 sentences uttered by a male speaker MHT and 503 sentences uttered by a female speaker FKN. We set interpolation ratio as  $(a_{\text{MHT}}, a_{\text{FKN}}) = (1, 0), (0.75, 0.25), (0.5, 0.5), (0.25, 0.75), (0, 1)$ .

By using speech parameter generation algorithm, 5 different types of synthesized speech were generated from 5 HMM sets. The MLSA filter was excited by pulse train or white noise generated according to pitch



**Figure 5. Results of ABX listening tests.**



**Figure 6. Subjective distance between samples.**

contours. To observe only a change of spectrum, pitch contours, which were linearly interpolated pitches extracted from MHT's and FKN's natural speech at a ratio of 1 : 1, was fixed.

The following audio files contain examples used in the experiment. [SOUND A1015S01.WAV], [SOUND A1015S02.WAV], [SOUND A1015S03.WAV], [SOUND A1015S04.WAV], [SOUND A1015S05.WAV] contains speech for  $(a_{\text{MHT}}, a_{\text{FKN}}) = (1, 0), (0.75, 0.25), (0.5, 0.5), (0.25, 0.75), (0, 1)$ .

##### 4.1. Generated Spectra

Fig. 4 shows spectra of a Japanese sentence "/n-i-m-o-ts-u-w-a/" generated from the triphone HMM sets. From this figure, it can be seen that spectra change smoothly from speaker MHT to speaker FKN by changing the interpolation ratio.

##### 4.2. ABX Listening Tests

Subjects were 8 males. In these tests, 4 sentences, which were different from training data, were synthesized and tested. Stimuli A and B were either MHT's or FKN's synthesized speech. Stimulus X was either 5 utterance synthesized with different interpolation ratio. Subjects listened this 5 utterance at random and were asked to select either A or B as being the closest.

Fig. 5 shows the experimental results. Horizontal axis represents the rate that speech samples from interpolated HMM sets were judged to be closer to that from MHT's HMMs. The figure shows the synthesized speech judged to be closer to MHT when  $a_{\text{MHT}} > a_{\text{FKN}}$ , and vice versa. This result suggests that the interpolated HMM sets maintains the characteristics of the representative speakers.

##### 4.3. Experiments of Similarity

In these tests, 2 sentences, which were different from training data, were synthesized and tested. Subjects were 8 males. Stimuli consisted of 2 samples in 5 utterance which were different interpolation ratio. Subjects were asked to rate the similarity of each pair into five categories ranging from "similar" to "dissimilar".

From the results, We placed each sample in a space according to the similarities between the samples

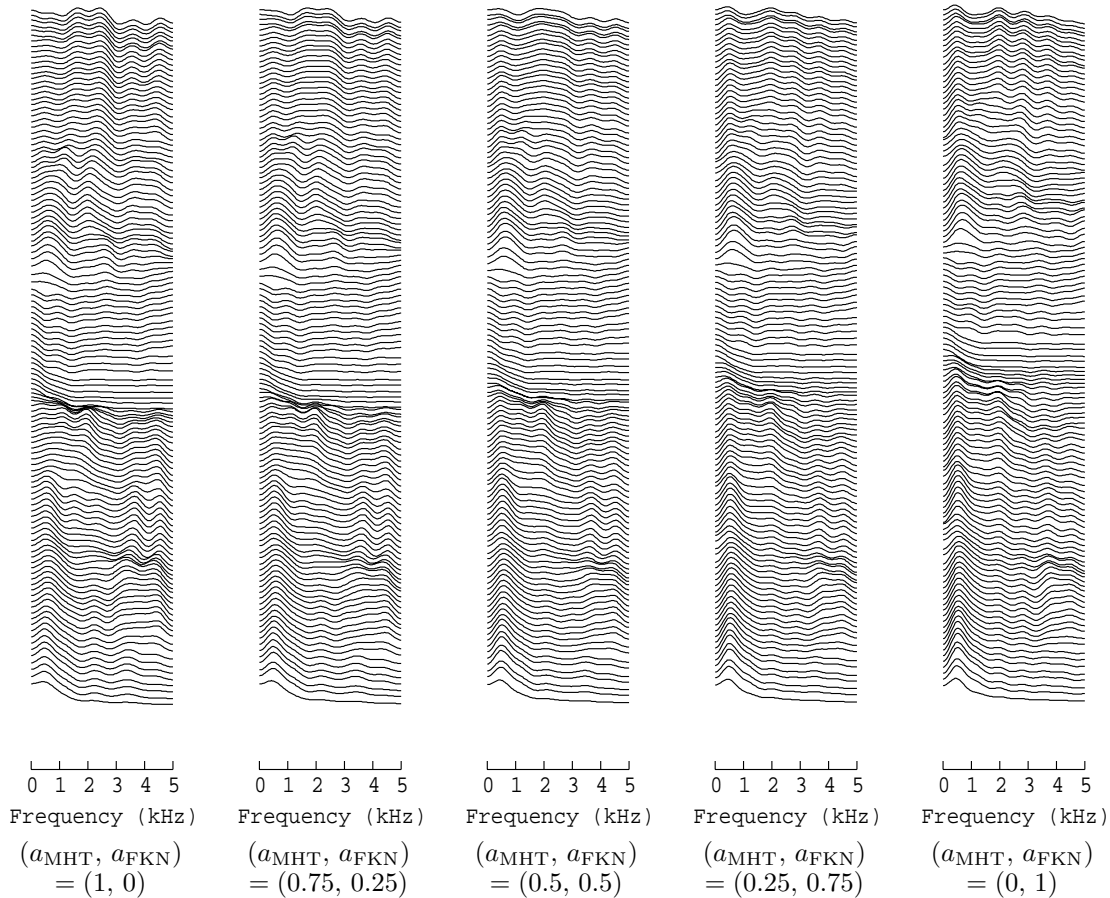


Figure 4. Generated spectra of the sentence “/n-i-m-o-t-s-u-w-a/”.

by using Hayashi’s fourth method of quantification. Fig. 6 shows the relative similarity-distance between stimuli. From the figure, it can be seen that synthesized speech changed smoothly from speaker MHT’s to speaker FKN’s by changing the interpolation ratio.

## 5. CONCLUSION

In this paper, we described an approach to voice characteristics conversion for a HMM-based text-to-speech synthesis system by interpolating HMMs of representative speakers. From the results of experiments, we have seen that the characteristics of synthesized speech from interpolated HMM set changed smoothly from one male speaker’s to the other female speaker’s by changing the interpolation ratio. The subjective experiments for the interpolation of multiple speakers and the investigation of selection method of the representative speakers are the future problems. We expect that the emotion (e.g. anger, sadness, joy) interpolation will be possible by replacing HMMs of representative speakers with HMMs of representative emotions.

## ACKNOWLEDGEMENT

This work was partially supported by Saneyoshi scholarship foundation and the Okawa foundation for information and telecommunications.

## REFERENCES

- [1] K. Tokuda, T. Kobayashi and S. Imai, “Speech parameter generation from HMM using dynamic features”, Proc. of ICASSP, 1, pp.660–663, May. 1995.
- [2] 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, 1, pp.757–760, Sep. 1995.
- [3] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, “Speech synthesis from HMMs using dynamic features”, Proc. of ICASSP, 1, pp.389–392, May. 1996.
- [4] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, “HMM-based speech synthesis with various voice characteristics”, Joint ASA/ASJ Meeting, Dec. 1996.
- [5] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, “Voice characteristics conversion for HMM-based speech synthesis system”, Proc. of ICASSP, Apr. 1997.
- [6] N. Iwahashi and Y. Sagisaka, “Speech Spectrum Conversion Based on Speaker Interpolation and Multi-functional Representation with Weighting by Radial Basis Function Networks”, Speech Communication, 16, pp.139-151, 1995.
- [7] S. Imai, “Cepstral analysis synthesis on the mel frequency scale”, Proc. of ICASSP, pp.93-96, 1983.
- [8] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech”, Proc. ICASSP-92, pp.1-137-1-140, 1992.
- [9] C. Hayashi, “Recent theoretical and methodological developments in multidimensional scaling and its related method in Japan”, Behaviormetrika, No.18, 1095, 1985.