SPEAKER INTERPOLATION IN HMM-BASED SPEECH SYNTHESIS SYSTEM

Takayoshi Yoshimura[1], *Takashi Masuko*[2], *Keiichi Tokuda*[1], *Takao Kobayashi*[2] *and Tadashi Kitamura*[1]

[1]Department of Computer Science, Nagoya Institute of Technology, Nagoya 466, Japan
[2]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 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 results of subjective experiments show that we can gradually change the characteristics of synthesized speech 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 required. However, it is difficult to collect, segment, and store them.

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

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. The idea of using speaker interpolation for voice conversion is similar to that of [6]. However, in the proposed method, each speech unit is modeled by an HMM, accordingly mathematically-well-defined statistical distance measures can be used. Listening tests show that the proposed algorithm successfully interpolates between representative speakers 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 text-to-speech synthesis system [3] 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 text-to-speech synthesis system based on speaker interpolation is shown in Fig. 1, which is almost equivalent to the previously proposed system except that multiple speaker's HMM sets are trained and a new speaker's HMM set is generated by interpolation between them. The procedure can be summarized as follows:

1. Training representative HMM sets
   (a) Select several representative speakers S_1, S_2, $\ldots$, S_N from speech database.
   (b) Obtain mel-cepstral coefficients from speech of the representative speakers by mel-cepstral analysis [8].
   (c) Train phoneme HMM sets Λ_1, Λ_2, $\ldots$, Λ_N for S_1, S_2, $\ldots$, S_N, respectively, using mel-cepstral coefficients, and their deltas and delta-deltas.
2. Interpolation between representative HMM sets
   (a) Generate a new phoneme HMM set Λ by interpolating between the representative speakers' phoneme HMM sets Λ_1, Λ_2, $\ldots$, Λ_N with an arbitrary interpolation ratio $a_1, a_2, \ldots, a_N$ based on a method described in the next section.
3. Speech synthesis from interpolated HMM
   (a) Convert the text to be synthesized into a phoneme sequence, and concatenate the interpolated phoneme HMMs according to the phoneme sequence.
   (b) Generate mel-cepstral coefficients from the sentence HMM by using speech parameter generation algorithm [1], [2].
   (c) Synthesize speech from the generated mel-cepstral coefficients by using the MLSA (Mel Log Spectral Approximation) filter [7].

In the mel-cepstral analysis [8], the mel-cepstral coefficients, i.e., frequency-transformed cepstral coefficients, are determined by maximizing $P(\mathbf{x} \,|\, \mathbf{c})$, where

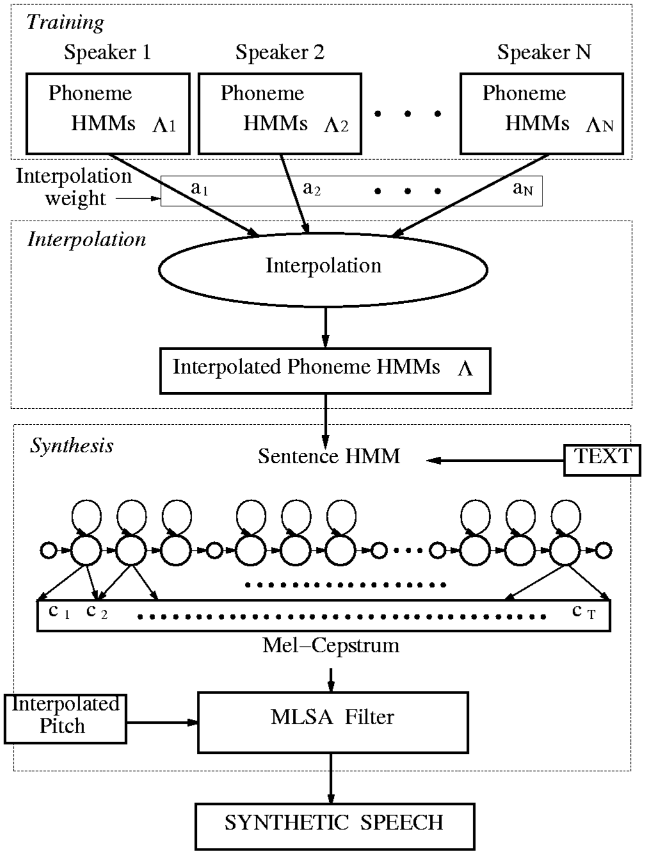

Figure 1. Block diagram of speech synthesis system.

$\mathbf{c}$ is the mel-cepstral coefficients, and $\mathbf{x}$ is the input speech sequence assumed to be zero-mean Gaussian. The transfer function defined by the mel-cepstral coefficients is realized by the MLSA filter [7] in which the filter coefficients are given by the mel-cepstral coefficients directly. By exciting the MLSA filter by pulse train or white noise generated according to a pitch contour, we can synthesize speech from the mel-cepstral coefficients with a small computational cost.

3. SPEAKER INTERPOLATION

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

$$I(\lambda, \lambda_k) = \int_{-\infty}^{\infty} P(\mathbf{O} \mid \lambda) \log \frac{P(\mathbf{O} \mid \lambda)}{P(\mathbf{O} \mid \lambda_k)} \, d\mathbf{O}, \tag{1}$$

where $\mathbf{O}$ is the speech parameter sequence. Then, for given HMMs $\lambda_1, \lambda_2, \ldots, \lambda_N$ and weights $a_1, a_2, \ldots, a_N$, consider a problem to obtain HMM λ which minimizes a cost function

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

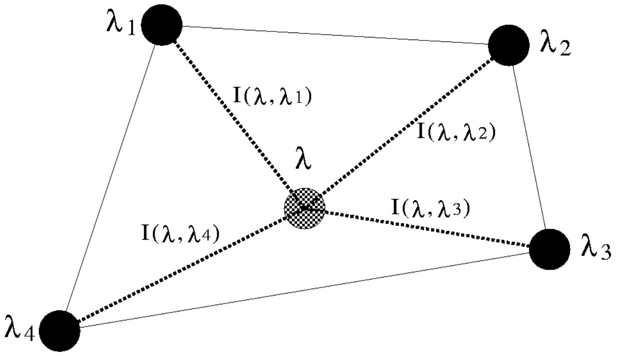

Figure 2. A space of speaker individuality modeled by HMMs.

We assume that representative speaker's HMMs have the same topology[1]. Under this assumption, interpolation between HMMs is equivalent to interpolation between output probability densities of 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, $p_k(\mathbf{o}) = \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}_k, \mathbf{U}_k)$, $\quad k = 1, 2, \ldots, N$, where $\boldsymbol{\mu}_k$ and $\mathbf{U}_k$ denote mean vector and covariance matrix, respectively, and $\mathbf{o}$ is the speech parameter vector. Consequently, we can determine the interpolated pdf $p(\mathbf{o}) = \mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \mathbf{U})$ by minimizing

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

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

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

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

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

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

respectively.

Fig. 3 shows Gaussian distributions generated by interpolation between Gaussian distributions p_1 and p_2 with two-dimensional diagonal covariances. In this figure, each ellipse represents the contour corresponding to the standard deviation of a Gaussian distribution, and each dot represents the mean vector of the distribution.

[1] Distributions could be tied.

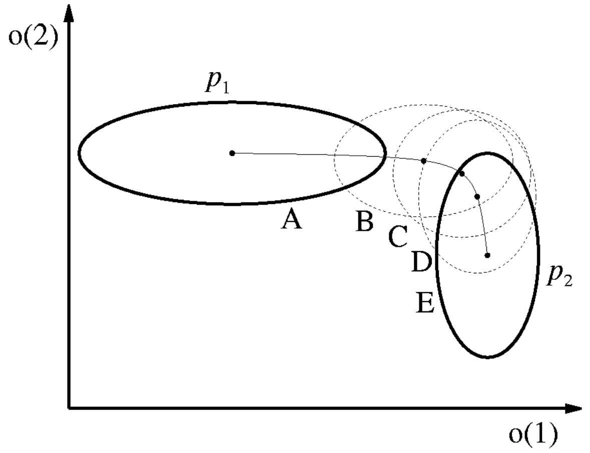

Figure 3. Interpolation between two Gaussian distributions p_1 and p_2 with interpolation ratios A : $(a_1, a_2) = (1, 0)$, B : $(a_1, a_2) = (0.75, 0.25)$, C : $(a_1, a_2) = (0.5, 0.5)$, D : $(a_1, a_2) = (0.25, 0.75)$, E : $(a_1, a_2) = (0, 1)$.

From the figure, it is seen that we can interpolate between two distributions appropriately in the sense that the interpolated distribution p reflects the statistical information , i.e., covariances of p_1 and p_2.

4. EXPERIMENTS

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

We used phonetically balanced 503 sentences from ATR Japanese speech database for training. Speech signals were sampled at 10kHz and 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 14 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 distributions. Decision-tree based model clustering was applied to each set of triphone models, and the resultant set of tied triphone models had approximately 2,800 distributions.

We trained two HMM sets using speech data from a male speaker MHT and a female speaker FKN, respectively. By using the speech parameter generation algorithm, five different types of speech were synthesized from five HMM sets obtained by setting the interpolation ratio as $(a_{\mathrm{MHT}}, a_{\mathrm{FKN}}) = (1, 0)$, (0.75, 0.25), (0.5, 0.5), (0.25, 0.75), (0, 1). The MLSA filter was excited by pulse train or white noise generated according to pitch contours. Pitch contours were extracted from MHT's and FKN's natural speech, and linearly interpolated at a ratio of 1 : 1 with Viterbi alignment. To observe only a change of spectrum, pitch contour was fixed for each sentence.

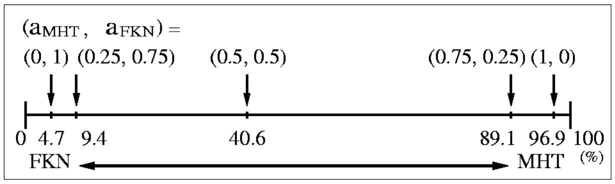

Figure 5. Results of ABX listening tests.

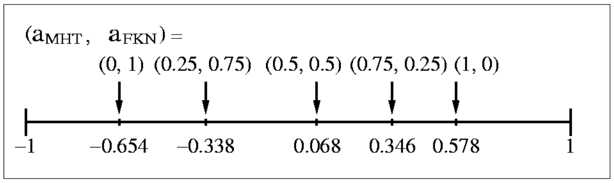

Figure 6. Subjective distance between samples.

The following audio files contain synthesized speech used in the experiments:

[SOUND A1015S01.WAV],
[SOUND A1015S02.WAV],
[SOUND A1015S03.WAV],
[SOUND A1015S04.WAV],
[SOUND A1015S05.WAV]

for $(a_{\mathrm{MHT}}, a_{\mathrm{FKN}}) = (1, 0)$, (0.75, 0.25), (0.5, 0.5), (0.25, 0.75), (0, 1), respectively.

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 the figure, it can be seen that spectra change smoothly from speaker MHT's to speaker FKN's according to the interpolation ratio.

4.2. ABX Listening Tests

Subjects were eight males. In this experiment, four sentences, which were not included in the training data, were synthesized and tested. Stimuli A and B were either MHT's or FKN's synthesized speech. Stimulus X was either of the five utterances synthesized with different interpolation ratios. Subjects listened this five utterances 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 HMM set. The figure shows the synthesized speech judged to be closer to MHT's when $a_{\mathrm{MHT}} > a_{\mathrm{FKN}}$, and vice versa. This result suggests that the interpolated HMM sets maintain the characteristics of the representative speakers.

4.3. Experiment of Similarity

In this experiment, two sentences, which were not included in the training data, were synthesized. Subjects were eight males. Stimuli consisted of two samples in five utterances which were synthesized with different interpolation ratios. Subjects were asked to rate the similarity of each pair into five categories ranging from "similar" to "dissimilar". From the results, we placed

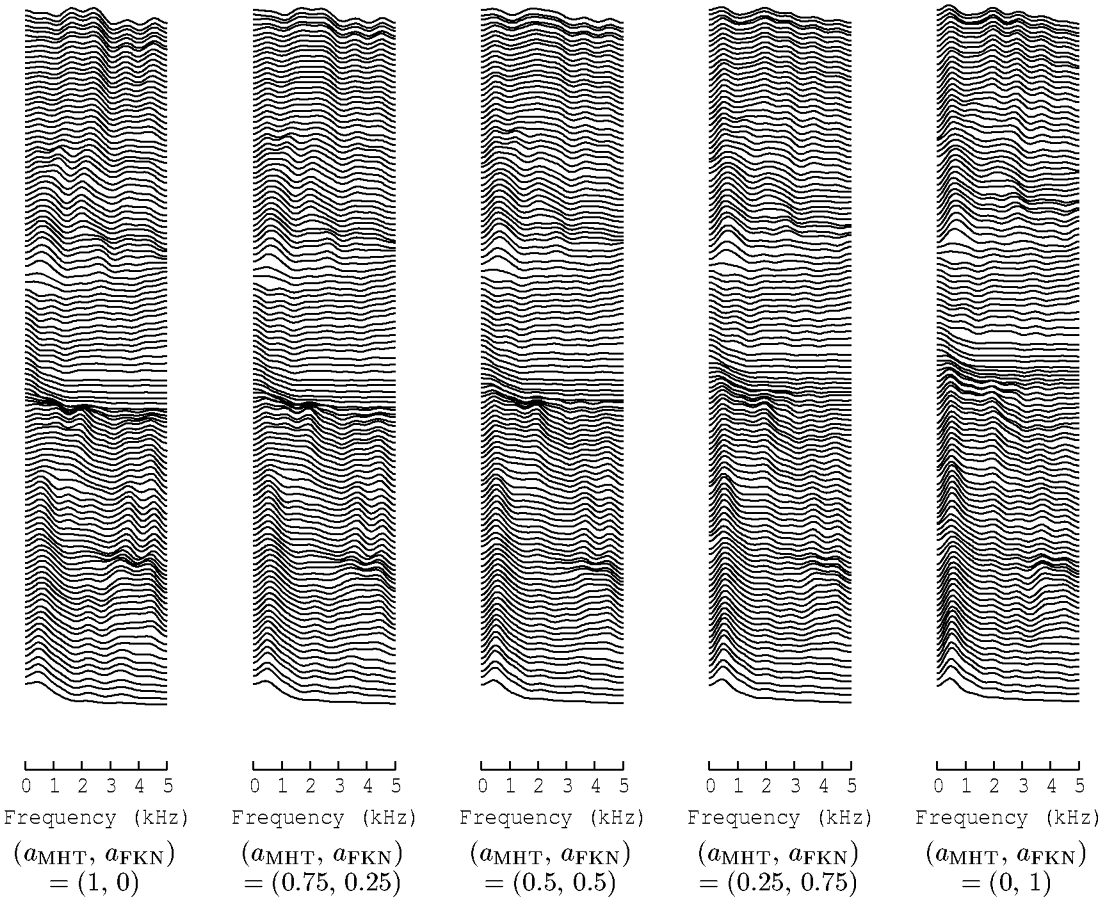

Figure 4. Generated spectra for a sentence "/n-i-m-o-ts-u-w-a/".

each sample in a space according to the similarities between the samples 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 according to the interpolation ratio.

5. CONCLUSION

In this paper, we described an approach to voice characteristics conversion for an HMM-based text-to-speech synthesis system by interpolation between HMMs of representative speakers. From the results of experiments, we have seen that the characteristics of synthesized speech from interpolated HMM set change smoothly from one male speaker's to the other female speaker's according to the interpolation ratio. The subjective experiments for the interpolation between 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

The authors would like to thank Prof. Satoshi Imai for useful discussions. 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, " Proc. of ASA/ASJ Third Joint Meeting, pp.1043-1046, Dec. 1996.

[5] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," Proc. of ICASSP, pp.1611-1614, 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. of ICASSP, pp.I-137–I-140, 1992.

[9] C. Hayashi, "Recent theoretical and methodological developments in multidimensional scaling and its related method in Japan," Behaviormetrika, No.18, 1095, 1985.