VOICE CHARACTERISTICS CONVERSION FOR HMM-BASED SPEECH SYNTHESIS SYSTEM

Takashi Masuko†, *Keiichi Tokuda*††, *Takao Kobayashi*† *and Satoshi Imai*†

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, Yokohama, 226 Japan
††Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466 Japan

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

ABSTRACT

In this paper, we describe an approach to voice characteristics conversion for an HMM-based text-to-speech synthesis system. Since this speech synthesis system uses phoneme HMMs as speech units, voice characteristics conversion is achieved by changing HMM parameters appropriately. To transform the voice characteristics of synthesized speech to the target speaker, we applied MAP/VFS algorithm to the phoneme HMMs. Using 5 or 8 sentences as adaptation data, speech samples synthesized from a set of adapted tied triphone HMMs, which have approximately 2,000 distributions, are judged to be closer to the target speaker by 79.7% or 90.6%, respectively, in an ABX listening test.

1. INTRODUCTION

In general, it is desirable that speech synthesis systems have the ability to synthesize speech with arbitrary voice characteristics. Furthermore, for speech translation systems, considering that these systems will be used by many speakers simultaneously, it is necessary to reproduce speaker's voice characteristics to make listeners possible to tell who is the speaker of the translated speech. From these points of view, we describe an approach to voice characteristics conversion for an HMM-based speech synthesis system.

We have proposed an algorithm for speech parameter generation from HMMs, and shown that using dynamic features we can generate a smoothly varying speech parameter sequence according to the statistical information of static and dynamic features modeled by HMMs [1][2]. We have also proposed an HMM-based text-to-speech synthesis system using this algorithm [3]. Since this system uses phoneme HMMs as speech units, voice characteristics conversion is achieved by transforming HMM parameters appropriately. In this paper, we apply MAP/VFS algorithm [4]-[7], one of the successful and widely used speaker adaptation techniques, to the synthesis system to transform voice characteristics from one speaker to another.

2. AN OVERVIEW OF THE HMM-BASED TEXT-TO-SPEECH SYNTHESIS SYSTEM

A block diagram of the HMM-based speech synthesis system [3] is shown in Fig. 1. The HMM-based speech synthesis system consists of two parts; training part and synthesis part. First, in the training part, mel-cepstral coefficients are obtained from speech database by mel-cepstral analysis [8]. Dynamic features, i.e., delta and delta-delta mel-cepstral coefficients, are calculated from mel-cepstral coefficients. Then phoneme HMMs are trained using mel-cepstral coefficients and their deltas and delta-deltas.

In the synthesis part, an arbitrary text to be synthesized is transformed into a phoneme sequence. We construct a sentence HMM, which represents the whole text to be synthesized, by concatenating phoneme HMMs according to this phoneme sequence. From the sentence HMM, speech parameter sequence is generated using the algorithm described in the next section. By using the MLSA (Mel Log Spectral Approximation) filter [9], speech is synthesized from the generated mel-cepstral coefficients directly.

2.1. Speech Parameter Generation from Continuous HMMs

In this paper, we assume that phoneme HMMs are left-to-right models with single Gaussian output distribution for convenience.

Let $\boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ be the vector sequence of speech parameter. We assume that the vector of speech parameter $\boldsymbol{o}_t$ at frame t consists of the static feature vector $\boldsymbol{c}_t$ and the dynamic feature vectors $\Delta\boldsymbol{c}_t$, $\Delta^2\boldsymbol{c}_t$, that is, $\boldsymbol{o}'_t = [\boldsymbol{c}'_t, \Delta\boldsymbol{c}'_t, \Delta^2\boldsymbol{c}'_t]'$. Dynamic features are

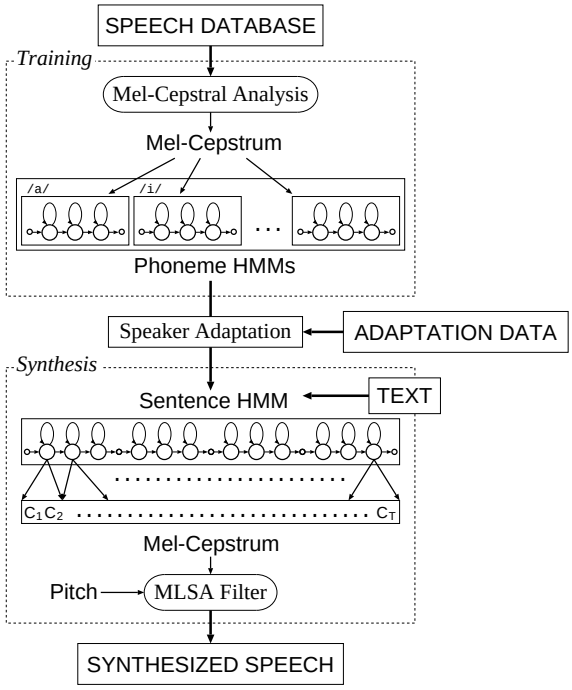


Figure 1: Block diagram of speech synthesis system.

calculated from the static features as follows,

$$\Delta \boldsymbol{c}_t = \frac{\sum_{\tau=-L_1}^{L_1} \tau \boldsymbol{c}_{t+\tau}}{\sum_{\tau=-L_1}^{L_1} \tau^2}, \quad (1)$$

$$\Delta^2 \boldsymbol{c}_t = \frac{\sum_{\tau=-L_2}^{L_2} \tau \Delta \boldsymbol{c}_{t+\tau}}{\sum_{\tau=-L_2}^{L_2} \tau^2}. \quad (2)$$

For a given continuous HMM λ, the output speech parameter sequence $\boldsymbol{c} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T\}$ is obtained by maximizing $P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T)$ with respect to state sequence $\boldsymbol{q}$ and parameter sequence $\boldsymbol{c}$ under the constraints of equations (1) and (2), that is, dynamic features and their output probabilities are determined by the generated speech parameter sequence $\boldsymbol{c}$ [1][2].

3. SPEAKER ADAPTATION USING A MAP/VFS ALGORITHM

To convert voice characteristics of synthesized speech to those of the target speaker we adapt the phoneme HMMs used as speech units (Fig. 1). By adapting phoneme HMMs, generated speech parameters become closer to the target speaker. As a result, voice characteristics of synthesized speech become closer to the target speaker.

We applied MAP (maximum *a posteriori*) estimation [4][5] and VFS (Vector Field Smoothing) algorithm [6][7] to adapt HMMs. MAP estimation and VFS algorithm are carried out sequentially for each adaptation data. These algorithms are described briefly in the following.

3.1. MAP Estimation

A feature of MAP estimation [4][5] is that the training incorporates prior information of model parameters and new incoming training data into estimated parameters. This is effective to the problem of limited training data in adaptation.

Let $\boldsymbol{\mu}$ and $\boldsymbol{U}$ be the mean vector and covariance matrix of the output probability distribution, respectively. For given training samples $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the new mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{U}}$ are estimated as follows [5],

$$\hat{\boldsymbol{\mu}} = \frac{\alpha \boldsymbol{\mu} + \sum_{i=1}^{N} \boldsymbol{x}_i}{\alpha + N}, \quad (3)$$

$$\hat{\boldsymbol{U}} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i' - (\alpha + N)\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' + \frac{\beta}{\alpha}\boldsymbol{U} + \alpha \boldsymbol{\mu}\boldsymbol{\mu}'}{\beta + N}, \quad (4)$$

where α and β are constants which control the influence of prior information. According to [4], β is determined based on the number of samples. However, in this paper, we determined β experimentally in a similar manner to [5]. We apply the MAP estimation to initial models as follows; (1) for a given adaptation data, HMMs are aligned using Viterbi algorithm to segment data samples, and (2) new means and covariances are estimated using samples associated to the states.

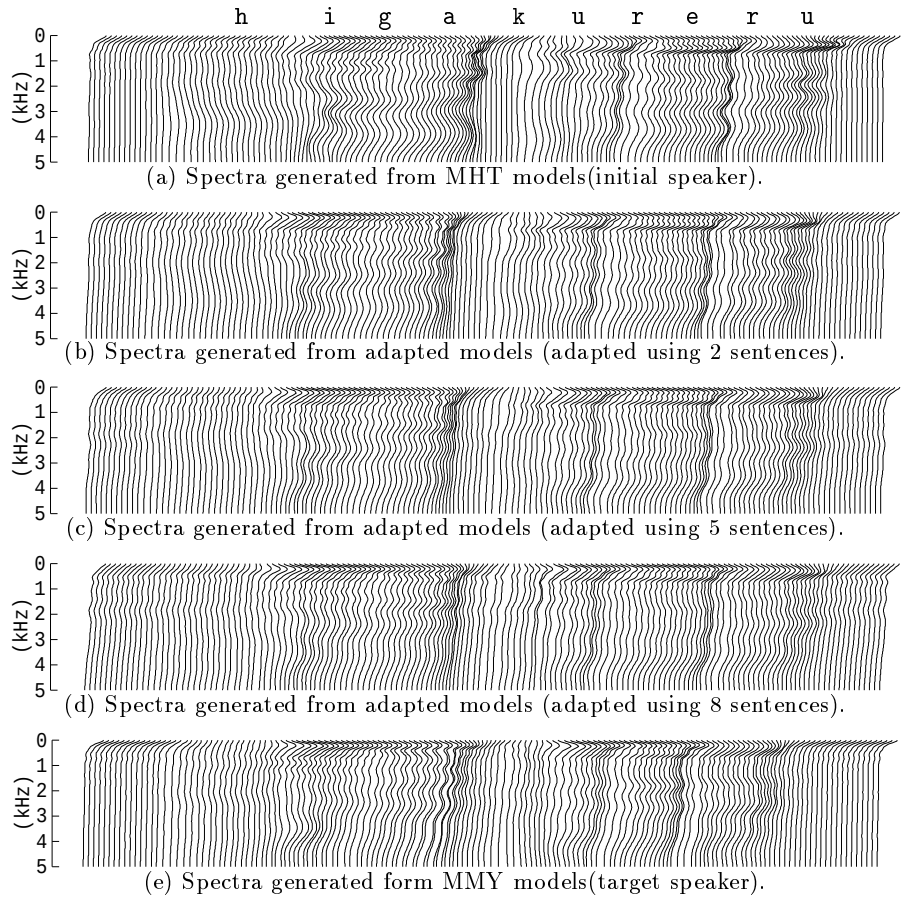
3.2. VFS algorithm

Since MAP estimation is carried out with very few data, there are many states which have no adaptation data and remain being untrained. Furthermore, MAP estimated parameters are unreliable because of insufficient training data. Therefore, we applied VFS algorithm [6][7] to interpolate new parameters of untrained distributions, and to smooth estimated parameters of MAP trained distributions.

3.2.1. Interpolation

Let G_K denotes the group of K nearest-neighbor distributions which are trained using MAP estimation. Interpolated mean vector and covariance of i-th untrained distribution, $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{U}}_i$, are calculated as follows,

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{j \in G_K} w_{ij}(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j)}{\sum_{j \in G_K} w_{ij}} + \boldsymbol{\mu}_i, \quad (5)$$



(a) Spectra generated from MHT models(initial speaker).

(b) Spectra generated from adapted models (adapted using 2 sentences).

(c) Spectra generated from adapted models (adapted using 5 sentences).

(d) Spectra generated from adapted models (adapted using 8 sentences).

(e) Spectra generated form MMY models(target speaker).

Figure 2: Generated spectra of the sentence "/h-i-g-a-k-u-r-e-r-u/"

$$\tilde{\boldsymbol{U}}_i = \frac{\sum_{j \in G_K} w_{ij} (\hat{\boldsymbol{U}}_j \boldsymbol{U}_j^{-1})}{\sum_{j \in G_K} w_{ij}} \boldsymbol{U}_i, \tag{6}$$

$$w_{ij} = \exp(-d_{ij}^2 / s), \tag{7}$$

where covariance matrices are assumed to be diagonal. In (5) and (6), $\boldsymbol{\mu}_j$ and $\boldsymbol{U}_j$ represent the initial mean vector and covariance matrix of j-th distribution, $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{U}}_j$ represent the trained mean vector and covariance matrix of j-th distribution using MAP estimation, respectively. We denote weighting for the neighboring transfer vector and matrix as w_{ij}, which is calculated by (7) where d_{ij} is Euclidean distance between mean vectors of i-th and j-th distributions and s is the smoothing parameter for adjusting the influence of neighboring distributions.

3.2.2. Smoothing

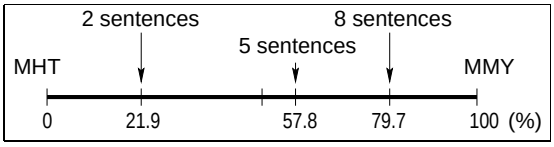
MAP estimated parameters are smoothed to increase those reliability. Smoothed mean vector and covariance matrix of i-th MAP trained distribution, $\bar{\boldsymbol{\mu}}_i$ and $\bar{\boldsymbol{U}}_i$, are calculated in the same manner to interpolation as follows,

$$\bar{\boldsymbol{\mu}}_i = \frac{(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) + \sum_{j \in G_K} w_{ij} (\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j)}{1 + \sum_{j \in G_K} w_{ij}} + \boldsymbol{\mu}_i, \tag{8}$$

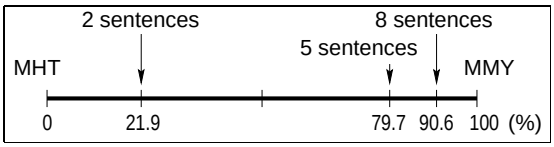
$$\bar{\boldsymbol{U}}_i = \frac{(\hat{\boldsymbol{U}}_i \boldsymbol{U}_i^{-1}) + \sum_{j \in G_K} w_{ij} (\hat{\boldsymbol{U}}_j \boldsymbol{U}_j^{-1})}{1 + \sum_{j \in G_K} w_{ij}} \boldsymbol{U}_i. \tag{9}$$

4. EXPERIMENTS

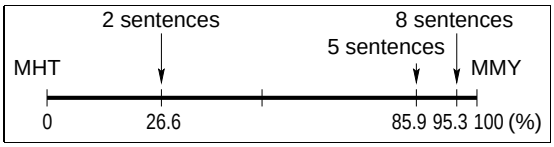
The effectiveness of speaker adaptation for the HMM-based speech synthesis system was evaluated through ABX listening tests. We used phonetically balanced 503 sentences from ATR Japanese speech database for training and adaptation. Speech signals were windowed by a 25.6ms Blackman window with a 5ms shift, then mel-cepstral coefficients were obtained by mel-cepstral analysis [8]. The feature vectors consisted of 16 mel-cepstral coefficients including the 0-th coefficient, and their delta and delta-delta coefficients. We set $L_1 = L_2 = 1$ in equations (1), (2).



(a) Triphone HMMs.



(b) Tied triphone HMMs (approximately 2,000 states).



(c) Tied triphone HMMs (approximately 1,200 states).

Figure 3: Results of ABX listening tests.

We used 5-state left-to-right triphone models with single Gaussian diagonal output distribution. In the database, there exist 3,518 distinct triphones. We also used two sets of tied triphone models which have approximately 2,000 and 1,200 distributions respectively (the set of triphone models have more than 17,000 distributions). Quality of synthesized speech from tied triphone HMM sets is comparable to that from triphone HMM sets (see [3]). HMMs were trained using 503 sentences uttered by a male speaker MHT (we denote these models as MHT models) and adapted using first 2, 5, and 8 sentences uttered by another male speaker MMY (adapted models). We also trained HMMs using 503 sentences uttered by MMY (MMY models) to compare with the adapted models. We set the parameters for MAP estimation as $\alpha = 15$ and $\beta = 50$, and set the number of nearest-neighbor $K = 5$ and smoothing factor $s = 10$ for VFS algorithm.

4.1. Generated Spectra

Fig. 2 shows spectra of a Japanese sentence "`/h-i-g-a-k-u-r-e-r-u/`" generated from tied triphone HMM sets which have approximately 2,000 distributions. From Fig. 2, it can be seen that the spectra generated from adapted models are getting closer to those from MMY models as the number of adaptation data increases.

4.2. ABX Listening Tests

ABX listening tests were conducted for speech synthesized from adapted models. Subjects were 7 males and 1 female. In these tests, 4 sentences, which were different from adaptation data, were synthesized and tested, where pitch contours were extracted from natural MHT's speech.

Fig. 3 shows the results of the listening tests. Horizontal axes represent the rates that speech samples from adapted models were judged to be closer to that from target models. Fig. 3(a) is the result for triphone models, and (b) and (c) are the results for tied triphone models. From these results, it can be seen that 5 or 8 sentences are sufficient to adapt HMMs, and that the required amount of data to adapt models decreases as the number of distributions decreases.

5. CONCLUSION

In this paper, we described an approach to voice characteristics conversion for the HMM-based text-to-speech synthesis system using MAP/VFS algorithm, and showed that we can easily vary voice characteristics by adapting HMM parameters to the target speaker. From the results of experiments, we have seen that a few sentences are sufficient to adapt HMMs from one speaker to another.

6. REFERENCES

[1] K. Tokuda, T. Kobayashi and S. Imai, "Speech Parameter Generation From HMM Using Dynamic Features," in *Proc. ICASSP-95*, pp.660-663, 1995.

[2] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and Satoshi Imai: "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features," in *Proc. EUROSPEECH-95*.

[3] T.Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech Synthesis Using HMMs with Dynamic Features," in *Proc. ICASSP-96*, pp.389-392, 1996.

[4] C. -H. Lee, C. H. Lin, and B. H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.39, no.4, pp.806-814, 1992.

[5] Y. Tsurumi and Seiichi Nakagawa, "An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum a Posteriori Probability Estimation," in *Proc. ICSLP-94*, S09-1.1, pp.431-434, 1994.

[6] J. Takahashi and S. Sagayama, "Vector-Field-Smoothed Bayesian Learning for Incremental Speaker Adaptation," in *Proc. ICASSP-95*, pp.696-699, 1995.

[7] J. Takahashi and S. Sagayama, "Variance Smoothing for Incremental Speaker Adaptation Method of MAP/VFS," in *Proc. Spring Meeting, Acostical Society of Japan*, 2-5-5, pp.39-40, March, 1995 (in Japanese).

[8] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, 1992, pp.I-137–I-140.

[9] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP-83*, pp.93-96, 1983.