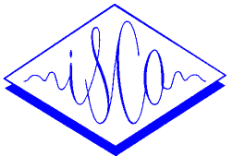


The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26–29, 1998

SPEAKER ADAPTATION FOR HMM-BASED SPEECH SYNTHESIS SYSTEM USING MLLR

Masatsune Tamura†, *Takashi Masuko*†, *Keiichi Tokuda*††, *and Takao Kobayashi*†

†Tokyo Institute of Technology, Yokohama, 226-8502 JAPAN
††Nagoya Institute of Technology, Nagoya, 466-8555 JAPAN

ABSTRACT

This paper describes a voice characteristics conversion technique for an HMM-based text-to-speech synthesis system. The system uses phoneme HMMs as the speech synthesis units, and voice characteristics conversion is achieved by changing HMM parameters appropriately. To transform the voice characteristics of synthetic speech to the target speaker, we apply an MLLR (Maximum Likelihood Linear Regression) technique, one of the speaker adaptation techniques, to the system. From the results of objective and subjective tests, it is shown that the characteristics of synthetic speech is close to target speaker's voice, and the speech generated from the adapted model set using 5 sentences has almost the same DMOS score as that from the speaker dependent model set.

1. INTRODUCTION

It is desirable that text-to-speech synthesis systems have the ability to synthesize speech with arbitrary voice characteristics. For instance, in the application of speech interpretation systems, translated speech should have the individual voice characteristics to identify the speaker. From this point of view, several techniques have been proposed to convert spectral parameters of input speech signals to have another speaker's characteristics [1][2].

We have proposed an algorithm for speech parameter generation from HMMs, and shown that we can generate a smoothly varying speech parameter sequence according to the statistical information of static and dynamic features modeled by HMMs [3]. Using this algorithm, we have also proposed an HMM-based text-to-speech synthesis system [4]. Since the system uses phoneme HMMs as the speech synthesis units, voice characteristics conversion can be achieved by transforming HMM parameters appropriately. In fact, we have shown that HMM parameters are adapted from one speaker to another using small amount of target speaker's speech data [5] by applying a MAP/VFS (Maximum *a Posteriori* / Vector Field Smoothing) algorithm [6][7].

In this paper, we propose an alternative approach to voice characteristics conversion by applying an MLLR (Maximum Likelihood Linear Regression) technique [8]-[10], one of the successful and widely used speaker adaptation techniques, to the HMM-based speech synthesis system. MAP/VFS uses multiple parameters to control the adaptation. As a result, it is not easy to determine an optimum set of the parameters for voice conversion. In contrast, MLLR requires only one parameter that represents the number of regression matrices.

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

2.1. HMM-based Speech Synthesis System

A block diagram of the HMM-based speech synthesis system [4][5] is shown in Figure 1. The system consists of three stages; the training stage, the adaptation stage, and the synthesis stage.

First, in the training stage, mel-cepstral coefficients are obtained from speech database by mel-cepstral analysis [11][12]. Dynamic features, i.e., delta and delta-delta coefficients, are also calculated from the mel-cepstral coefficients. Then, for each phoneme, speaker independent HMM, we will refer to it as the initial phoneme HMM, is trained using mel-cepstral coefficients and their deltas and delta-deltas.

Secondly, in the adaptation stage, feature vectors are calculated from given adaptation data. Then, initial HMMs are transformed into the target speaker HMMs by applying a speaker adaptation technique. In this study, we adapt mean vectors and covariance matrices of output distributions of HMMs using the MLLR technique.

Finally, in the synthesis stage, an arbitrarily given 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 the phoneme sequence. From the obtained sentence HMM, mel-cepstral parameter sequence is generated using the algorithm [3] described briefly in the next section. By using the MLSA (Mel Log Spectrum Approximation) filter [12][13], speech is synthesized from the generated mel-cepstral coefficients.

2.2. Parameter Generation from HMM

Let $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ be the speech parameter vector sequence. We assume that the speech parameter vector $\boldsymbol{o}_t$ 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 calculated from the static features by

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

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

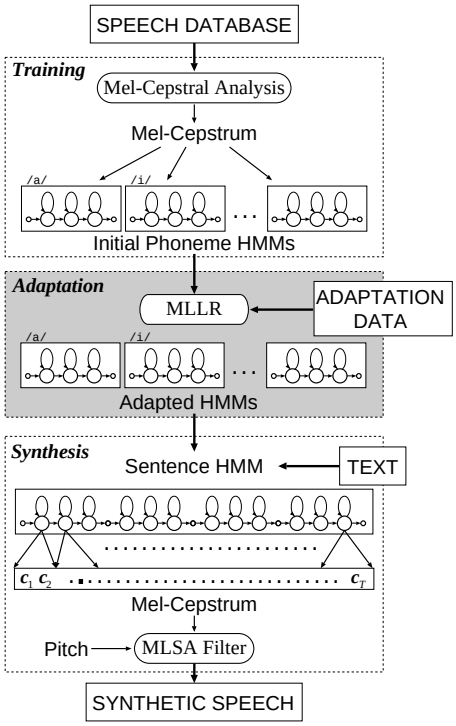


Figure 1: Block diagram of speech synthesis system.

For a given continuous HMM λ with single Gaussian output distributions, we can obtain a speech parameter vector sequence $\boldsymbol{O}$ that maximizes $P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)$ with respect to the state sequence $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$ and $\boldsymbol{C} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \ldots, \boldsymbol{c}_T']'$ under the constraints of eqs. (1) and (2) [3]. If the state sequence $\boldsymbol{Q}$ is explicitly known, the optimum parameter vector sequence is obtained by solving a set of linear equations which is derived from

$$\partial \log P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)/\partial \boldsymbol{C} = \boldsymbol{0}. \tag{3}$$

Without dynamic features, i.e., $\boldsymbol{o}_t = \boldsymbol{c}_t$, it is obvious that $P(\boldsymbol{Q}, \boldsymbol{O}|\lambda, T)$ is maximized when the parameter vector sequence is equal to the mean vector sequence which is determined independently of the covariances of the output distributions. On the other hand, by using delta parameters, the generated parameter vector reflects both means and covariances of the output distributions of a number of frames before and after the current frame.

3. SPEAKER ADAPTATION USING MLLR

To convert voice characteristics of synthetic speech to those of the target speaker, we adapt the initial phoneme HMMs to the target speaker HMMs (Figure 1). By doing this, parameters of output distributions of HMMs are modified to reflect the target speaker's voice characteristics. As a result, generated speech parameters become closer to those of the target speaker, and voice characteristics of synthetic speech also become closer to the target speaker.

In the previous work [5], we used a MAP/VFS algorithm for adaptation. The MAP/VFS algorithm requires multiple parameters to control the adaptation, such as weights for prior density, smoothing factor for VFS, and the number of PDFs to calculate transfer vectors [6][7]. Therefore it is not easy to determine a set of the parameters which provide the best performance in the HMM-based speech synthesis. In contrast, MLLR is based on maximum likelihood estimation, and the required parameter for adaptation is only the number of regression matrices. When the amount of adaptation data is small, we simply use one regression matrix.

The algorithms of the MLLR adaptation is described briefly in the following. We assume that HMMs have a single Gaussian output probability distribution for each state and the covariance matrix is diagonal.

3.1. Mean adaptation

Mean adaptation of MLLR is based on affine transformation. Let $\boldsymbol{\mu}_q$ and $\boldsymbol{U}_q$ be the mean vector and the covariance matrix of the output probability distribution of state q, respectively. For given training samples $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$, the new mean vector $\hat{\boldsymbol{\mu}}_q$ is estimated as follows [8]:

$$\hat{\boldsymbol{\mu}}_q = \boldsymbol{A}_q \boldsymbol{\mu}_q + \boldsymbol{b}_q = \boldsymbol{W}_q \boldsymbol{\xi}_q, \tag{4}$$

where $\boldsymbol{\xi}_q$ is an extended mean vector defined by $\boldsymbol{\xi}_q = [\ 1\ \boldsymbol{\mu}_q'\ ]'$, and $\boldsymbol{W}_q$ is the regression matrix for the mean vector. Define $\gamma_q(t)$ as the probability of generating $\boldsymbol{o}_t$ in state q at time t, given the observation sequence $\boldsymbol{O}$ and the model λ:

$$\gamma_q(t) = \frac{1}{P(\boldsymbol{O}|\lambda)} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P(\boldsymbol{O}, \theta_t = q|\lambda), \tag{5}$$

where $\boldsymbol{\Theta}$ is the set of all possible state sequences $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_T\}$ of length T. Then the regression matrix $\boldsymbol{W}_q$ is found by solving the following equation

$$\sum_{t=1}^{T} \sum_{r=1}^{R} \gamma_{q_r}(t) \boldsymbol{U}_{q_r}^{-1} \boldsymbol{o}_t \boldsymbol{\xi}_{q_r}' = \sum_{t=1}^{T} \sum_{r=1}^{R} \gamma_{q_r}(t) \boldsymbol{U}_{q_r}^{-1} \boldsymbol{W}_q \boldsymbol{\xi}_{q_r} \boldsymbol{\xi}_{q_r}', \tag{6}$$

where R is the number of states which share the regression matrix $\boldsymbol{W}_q$. When the amount of adaptation data is very small, we cannot calculate regression matrix for each state. In this case, we tie regression matrix among several states to adapt all distributions.

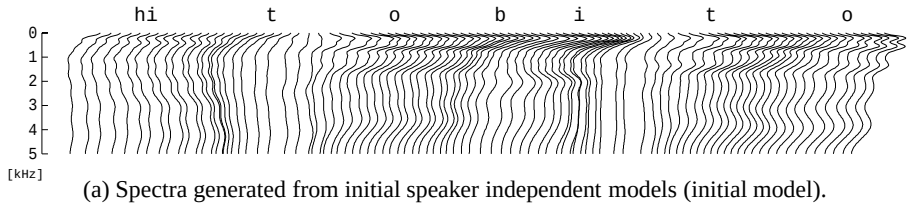
3.2. Variance adaptation

Covariance matrix adaptation is performed after the mean adaptation [9]. Let $\hat{\boldsymbol{U}}_q$ be the adapted covariance matrix

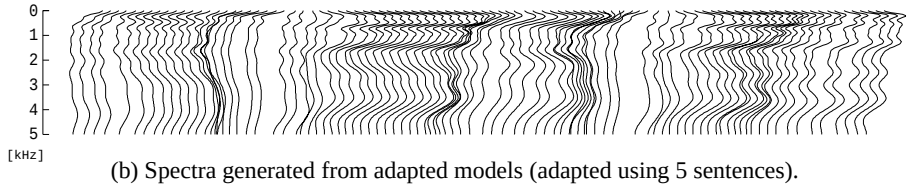
$$\hat{\boldsymbol{U}}_q = \boldsymbol{B}_q' \boldsymbol{H}_q \boldsymbol{B}_q, \tag{7}$$

where $\boldsymbol{C}_q$ is the Choleski factor of $\boldsymbol{U}_q^{-1}$ and $\boldsymbol{B}_q = \boldsymbol{C}_q^{-1}$. Regression matrix for the covariance matrix $\boldsymbol{H}_q$ is estimated by

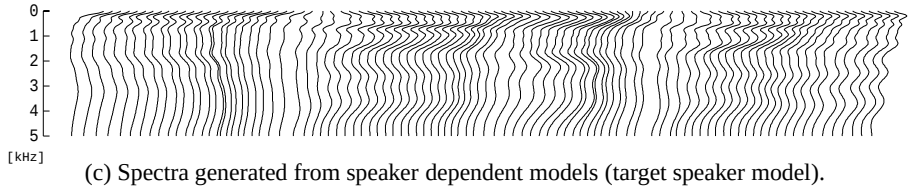
$$\boldsymbol{H}_q = \left(\sum_{r=1}^{R} \Big\{ \boldsymbol{C}_{q_r}' \Big[\sum_{t=1}^{T} \gamma_{q_r}(t) (\boldsymbol{o}(t) - \hat{\boldsymbol{\mu}}_{q_r}) (\boldsymbol{o}(t) - \hat{\boldsymbol{\mu}}_{q_r})' \Big] \boldsymbol{C}_{q_r} \Big\} \right) \Big/ \left(\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{q_r}(t) \right). \tag{8}$$



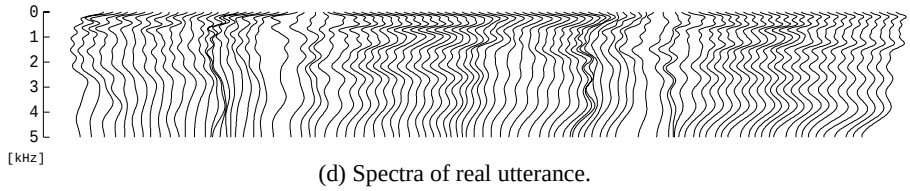
(a) Spectra generated from initial speaker independent models (initial model).



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



(c) Spectra generated from speaker dependent models (target speaker model).



(d) Spectra of real utterance.

Figure 2: Generated spectra of the sentence "/hi-t-o-b-i-t-o/"

In general, $\boldsymbol{H}_q$ becomes a full covariance matrix, even if the original covariance matrices were diagonal. By setting off-diagonal elements of $\boldsymbol{H}_q$ to zero, $\hat{\boldsymbol{U}}_q$ becomes a diagonal matrix.

4. EXPERIMENTS

The effectiveness of the speaker adaptation for the HMM-based speech synthesis system was evaluated through objective and subjective tests. Spectral distances between synthetic speech and real speech were calculated for the objective test, and DMOS listening tests were performed for the subjective test.

We used phonetically balanced sentences from ATR Japanese speech database for training and adaptation. The sampling frequency of the speech data is 10kHz. Speech signals were windowed by a 25.6ms Blackman window with a 5ms shift, then mel-cepstral coefficients were obtained by mel-cepstral analysis [11][12]. The feature vectors consisted of 16 mel-cepstral coefficients including the 0th coefficient, and their delta and delta-delta coefficients. We set $L_1 = L_2 = 1$ in eqs. (1), (2).

We used 5-state left-to-right triphone models with single diagonal Gaussian output distributions. Initial (speaker independent) HMMs were trained using 10 male speakers, 150 sentences for each speaker (we denote these models as SI models). The models were state clustered using a tree based clustering procedure[14]. The total number of states in SI models are 2,213. We set the target speaker to a male speaker MYI, who was not included in the speakers for training initial HMMs, in the database. Consequently, we adapted SI models to MYI using 1, 3, 5, 8 sentences uttered by MYI. It should be noted that the adaptation data was not contained in the training data. Since the amount of adaptation data was very small, we used one regression matrix, tied among all states, for MLLR adaptation. We also trained HMMs using 450 sentences uttered by MYI (SD models) to compare with the adapted models. We synthesized 3 sentences for evaluation tests. These sentences were contained in neither training nor adaptation data. We used pitch and duration information obtained from natural MYI's utterances for the speech synthesis.

4.1. Generated Spectra

Figure 2 shows spectra of a Japanese sentence "/hi-t-o-b-i-t-o/". The spectral envelopes generated from SI models (a), generated from adapted models using 5 sentence (b), generated from SD models (c), and real utterance (d) are shown. From Figure 2, it can be seen that the spectra generated from adapted models are closer to those from SD models than SI models.

4.2. Mel-Log-Spectral Distance

Figure 3 shows the mel-log-spectral distances between synthetic speech and real speech. The mel-log-spectral distance between $\boldsymbol{C} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \cdots, \boldsymbol{c}_T']'$ and $\overline{\boldsymbol{C}} = [\overline{\boldsymbol{c}}_1', \overline{\boldsymbol{c}}_2', \cdots, \overline{\boldsymbol{c}}_T']'$ is defined by

$$d(\boldsymbol{C}, \overline{\boldsymbol{C}}) = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{k=1}^{M} \left(c_t(k) - \bar{c}_t(k) \right)^2}, \qquad (9)$$

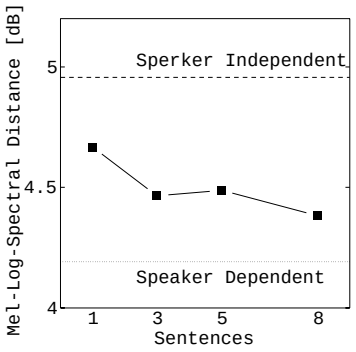


Figure 3: Mel-log-spectral distance.

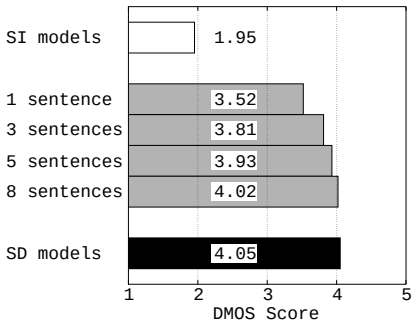


Figure 4: DMOS scores for synthetic speech.

where $c_t(k)$ denotes the *k*-th mel-cepstrum coefficient at time t. In Figure 3, the average of distances over 3 test sentences is shown. Horizontal axis represents the number of sentences used for adaptation data. Dashed line shows the distance of SI models and dotted line shows the distance of SD models. It is seen that distances of synthetic speech of adapted models are smaller than SI models, and the distance tends to get smaller as increasing the number of sentences.

4.3. DMOS Listening Tests

DMOS listening tests were conducted for evaluation of synthetic speech from adapted models. Subjects were 7 males. Figure 4 shows the results of the listening tests. DMOS scores of synthetic speech from SI models, adapted models using 1, 3, 5, 8 sentences, and SD models are shown. From the figure, it can be seen that the speech generated from adapted models using 5 or 8 sentences is judged to almost the same score of SD models.

We compared the MLLR adaptation with the MAP/VFS adaptation [6][7] in informal listening tests. In the MAP/VFS, we set control parameters to the values described in the original paper [6][7], in which the technique is used for speech recognition. The results showed that the MLLR adaptation provides higher performance than MAP/VFS adaptation. The system with MAP/VFS adaptation might achieve higher performance, if we find an optimum set of parameters. However, MLLR adaptation is efficient in the sense that synthetic speech are converted without parameter adjustment.

5. CONCLUSION

In this paper, we described an approach to voice characteristics conversion for the HMM-based text-to-speech synthesis system using MLLR algorithm. It has been shown that we can easily vary voice characteristics by adapting HMM parameters to the target speaker. From the experimental results, we have shown that a few sentences are sufficient to adapt HMMs.

The characteristics of speech is greatly depend on prosodic information, such as pitch contour and duration. Our future work is convert the characteristics of prosodic features in the same framework.

6. REFERENCES

1. Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol.6, no.2, pp.131–142, Mar. 1998.
2. M. Hashimoto and N. Higuchi, "Training data selection for voice conversion using speaker selection and vector field smoothing," *Proc. ICSLP-96*, pp.1397–1400, Oct. 1996.
3. K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features," *Proc. ICASSP-95*, pp.660–663, 1995.
4. T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis using HMMs with dynamic features," *Proc. ICASSP-96*, pp.389–392, 1996.
5. T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," *Proc. ICASSP-97*, pp.1611–1614, 1997.
6. Y. Tsurumi and S. Nakagawa, "An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation," *Proc. ICSLP-94*, S09-1.1, pp.431–434, 1994.
7. J. Takahashi, and S. Sagayama, "Vector - field - smoothed bayesian learning for fast and incremental speaker / telephone - channel adaptation," *Computer Speech and Language*, vol.11, no.2, pp.127–146, 1997.
8. C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continu ous density hidden Markov models," *Computer Speech and Language*, vol.9, no.2, pp.171–185, 1995.
9. M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol.10, no.4, pp.249–264, 1996.
10. M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol.12, no.2, pp.75–98, 1998.
11. K. Tokuda, T. Kobayashi, T. Fukada, H. Saito and S. Imai, "Spectral estimation of speech based on mel-cepstral representation," *Trans. IEICE*, vol. J74-A, pp.1240–1248, Aug. 1991. (in japanese).
12. T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP-92*, pp.137–140, 1992.
13. S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proc. ICASSP-83*, pp.93–96, 1983.
14. S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," *Proc. ARPA Human Language Technology Workshop*, pp.307–312, Mar. 1994.

AUDIO EXAMPLES

R53_01.wav

R53_02.wav

R53_03.wav

R53_04.wav