A TRAINING METHOD FOR AVERAGE VOICE MODEL BASED ON SHARED DECISION TREE CONTEXT CLUSTERING AND SPEAKER ADAPTIVE TRAINING

Junichi Yamagishi†, Takashi Masuko†, Keiichi Tokuda††, Takao Kobayashi†

†Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan

††Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan

Email: {Junichi.Yamagishi,masuko,Takao.Kobayashi}@ip.titech.ac.jp, tokuda@ics.nitech.ac.jp

ABSTRACT

This paper describes a new training method of average voice model for speech synthesis in which arbitrary speaker's voice is generated based on speaker adaptation. When the amount of training data is limited, the distributions of average voice model often have bias depending on speaker and/or gender and this will degrade the quality of synthetic speech. In the proposed method, to reduce the influence of speaker dependence, we incorporate a context clustering technique called shared decision tree context clustering and speaker adaptive training into the training procedure of average voice model. From the results of subjective tests, we show that the average voice model trained using the proposed method generates more natural sounding speech than the conventional average voice model. Moreover, it is shown that voice characteristics of synthetic speech generated from the adapted model using the proposed method are closer to the target speaker than the conventional method.

1. INTRODUCTION

A goal of text-to-speech (TTS) synthesis is to have an ability to generate natural sounding speech with arbitrary speaker's voice characteristics and various speaking styles. We believe that HMM-based speech synthesis using average voice model [1][2] is a promising approach to this problem. Average voice model is a set of speaker independent speech synthesis units trained using multi-speaker database for the HMM-based speech synthesis. To generate an arbitrarily given target speaker's voice, the average voice model is adapted to the target speaker using a speaker adaptation technique, such as MLLR (Maximum Likelihood Linear Regression) algorithm [3], and then HMM-based speech synthesis [4][5] is performed with the speaker adapted model. We have shown that a TTS system with the average voice model can generate synthetic speech which resembles the target speaker's voice by applying speaker adaptation technique based on MLLR using only a small amount of target speaker's speech data [1][2].

To obtain higher performance in the model adaptation to a wide variety of target speakers, the initial model of the adaptation, namely the average voice model, should not have any bias depending on speaker and/or gender. However, it would occur that the distributions of the average voice model have relatively large bias depending on speaker and/or gender included in the training speech database, especially when the amount of the training data is small. This will affect model adaptation performance and degrade the quality of synthetic speech. To overcome this problem, we proposed a technique for constructing a decision tree used for clustering the average voice model [6]. Using this technique, which we will call "shared decision tree context clustering (STC)" here, every node of the decision tree always has training data from all speakers included in the training speech database. As a result, each distribution of the average voice model reflects the statistics of all speakers. Moreover, it has been shown that the quality of the average voice improves by using this technique [6].

In this paper, we propose a new training method of average voice model for further reducing influence of speaker dependence and improving the quality of both average voice and synthetic speech of the given target speaker. In the proposing method, we incorporate speaker adaptive training (SAT) [7] as well as STC into the training procedure of the average voice model. Specifically, STC is used for clustering distributions of spectrum, pitch (F_0), and state duration, then SAT is used for re-estimation of parameters of spectrum and F_0. We show results of subjective evaluation of the proposing technique and also show its effectiveness.

2. TRAINING TECHNIQUE OF AVERAGE VOICE MODEL FOR SPEAKER ADAPTATION

2.1. Overview of Speech Synthesis from Average Voice

Speech synthesis system using average voice model is described in detail in [1][2]. The basic structure is the same as the HMM-based speech synthesis system [4][5] except that the average voice model is used as the set of synthesis units and speaker adaptation stage is added between the training and synthesis stages.

In the training stage, speaker independent phoneme HMMs are trained using multi-speaker 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) [8], respectively. To model variations of spectrum and F_0, phonetic 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 [9][10] 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.

2.2. Training of Average Voice Model

A block diagram of the training stage of the average voice model using the proposing technique is shown on the right side of Fig. 1.

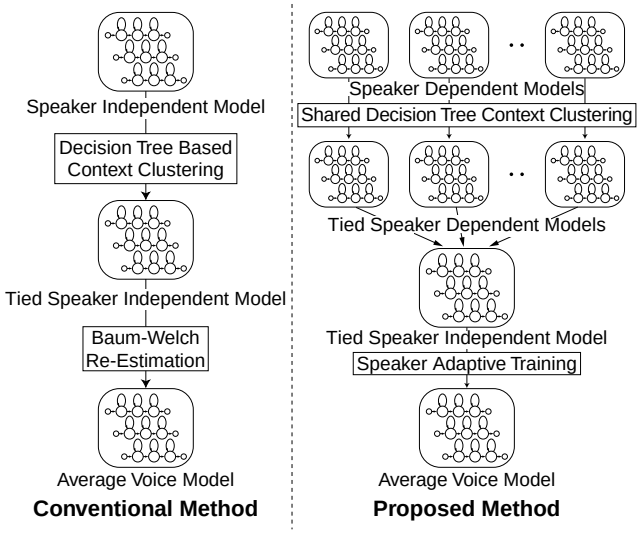


Fig. 1. A block diagram of the training stage of the average voice model.

First, context dependent models without context clustering are separately trained for respective speakers. Then, the decision tree, which we refer to as a shared decision tree, is constructed using an algorithm described in [6] from the speaker dependent models. All speaker dependent models are clustered using the shared decision tree. A Gaussian pdf of average voice model is obtained by combining all speakers' Gaussian pdfs at every node of the tree. After re-estimation of parameters of the average voice model using speaker adaptive training (SAT) [7] described in **2.4** with training data of all speakers, state duration distributions are obtained for each speaker. Finally, state duration distributions of the average voice model are obtained by applying the same clustering procedure.

2.3. Shared Decision Tree Context Clustering

In the shared decision tree context clustering (STC) [6], a speaker independent decision tree common to all speaker dependent models is constructed based on the minimum description length (MDL) criterion [10].

Let S_0 be the root node of a decision tree and $U(S_1, S_2, \ldots, S_M)$ be a model[1] defined for a leaf node set $\{S_1, S_2, \ldots, S_M\}$. A Gaussian pdf $\mathcal{N}_{im}$ of speaker i is assigned to each node S_m, and the set of Gaussian pdfs of each speaker i for the node set $\{S_1, S_2, \cdots, S_M\}$ is defined as $\lambda_i(S_1, S_2, \cdots, S_M) = \{\mathcal{N}_{i1}, \mathcal{N}_{i2}, \ldots, \mathcal{N}_{iM}\}$.

The log-likelihood of λ_i for the training data is given by

$$\begin{aligned} L(\lambda_i) &= \sum_{m=1}^{M} L(\mathcal{N}_{im}) \\ &= -\frac{1}{2}\sum_{m=1}^{M} \Gamma_{im}\left(K + K\log(2\pi) + \log|\boldsymbol{\Sigma}_{im}|\right), \end{aligned} \quad (1)$$

where Γ_{im} is the total state occupancy count at node S_m for speaker i, K is the dimensionality of the data vector, and $\boldsymbol{\Sigma}_{im}$ is the diagonal covariance matrix of the Gaussian pdf of speaker i at node S_m. Then, using (1), the description length of λ_i is given by

$$\begin{aligned} D(\lambda_i) &= -L(\lambda_i) + cKM\log W_i + C \\ &= \frac{1}{2}\sum_{m=1}^{M} \Gamma_{im}\left(K + K\log(2\pi) + \log|\boldsymbol{\Sigma}_{im}|\right) \\ &\quad + cKM\log W_i + C, \end{aligned} \quad (2)$$

where $W_i = \sum_{m=1}^{M} \Gamma_{im}$, and C is the code length required for choosing the model which is assumed here to be constant. Note that we introduce here a weight c for adjusting the model size.

We now define the description length for the model U as

$$\widehat{D}(U) = \sum_{i=1}^{I} D(\lambda_i), \quad (3)$$

where I is the total number of speakers. Suppose that node S_m of model U is split into two nodes S_{mqy} and S_{mqn} by applying a question q. Let U' be the model obtained by splitting S_m of model U by the question q. Then we define the difference between the description lengths after and before the splitting as follows:

$$\delta_m(q) = \widehat{D}(U') - \widehat{D}(U) \quad (4)$$

The procedure of construction of the shared decision tree is summarized as follows:

1. Define an initial model U as $U = \{S_0\}$.
2. Find the node $S_{m'}$ in model U and the question q' which minimizes $\delta_{m'}(q')$.
3. Terminate if $\delta_{m'}(q') > 0$.
4. Split the node $S_{m'}$ by the question q', and replace U by the resultant node set.
5. Go to step 2.

After the construction of the shared decision tree, we obtain Gaussian pdfs of the average voice model by combining Gaussian pdfs of speaker dependent models. The mean vector $\boldsymbol{\mu}_m$ and the covariance matrix $\boldsymbol{\Sigma}_m$ of the Gaussian pdf at node S_m are calculated as follows:

$$\boldsymbol{\mu}_m = \frac{\sum_{i=1}^{I} \Gamma_{im}\boldsymbol{\mu}_{im}}{\sum_{i=1}^{I} \Gamma_{im}}, \quad (5)$$

$$\boldsymbol{\Sigma}_m = \frac{\sum_{i=1}^{I} \Gamma_{im}\left(\boldsymbol{\Sigma}_{im} + \boldsymbol{\mu}_{im}\boldsymbol{\mu}_{im}^{\top}\right)}{\sum_{i=1}^{I} \Gamma_{im}} - \boldsymbol{\mu}_m\boldsymbol{\mu}_m^{\top}, \quad (6)$$

where $\cdot^{\top}$ denotes matrix transpose, and Γ_{im}, $\boldsymbol{\mu}_{im}$, and $\boldsymbol{\Sigma}_{im}$ are the state occupancy count, the mean vector, and the covariance matrix of the Gaussian pdf of speaker i at node S_m, respectively.

2.4. Speaker Adaptive Training

Here we incorporate the SAT paradigm [7] into the average voice model training. In SAT, speaker independent model is trained so that the resultant model of the MLLR-based speaker adaptation maximizes the likelihood for respective training speakers.

In MLLR-based speaker adaptation, the adapted mean vector $\hat{\boldsymbol{\mu}}_{im}$ of the state m of speaker i is estimated by

$$\hat{\boldsymbol{\mu}}_{im} = \boldsymbol{W}_i\boldsymbol{\xi}_m = \boldsymbol{A}_i\boldsymbol{\mu}_m + \boldsymbol{b}_i, \quad (7)$$

[1] Here a model represents a set of leaf node of decision tree. See [6].

Table 1. The number of distributions after clustering.

	NONE	SAT	STC	STC+SAT
Spec.	856	856	1251	1251
F_0	2742	2742	2217	2217
Dur.	1865	1487	2212	1821

where $\boldsymbol{\xi}_m = [1, \boldsymbol{\mu}_m^\top]^\top$, and $\boldsymbol{W}_i = [\boldsymbol{b}_i \boldsymbol{A}_i]$ is the regression matrix for the mean vector. In the SAT paradigm, the regression matrix $\boldsymbol{W}_i$ is re-estimated in accordance with a standard EM algorithm and the mean vectors and the covariance matrices of the Gaussian pdfs are re-estimated using the updated values of the regression matrices based on an extended EM algorithm. This re-estimation process is repeated until the convergence.

In the average voice model training, the maximum likelihood estimation of the mean vectors $\bar{\boldsymbol{\mu}}_m$ and the covariance matrices $\bar{\boldsymbol{\Sigma}}_m$ of the Gaussian pdfs in state m of speaker i for the training data $\boldsymbol{O}_i = \{\boldsymbol{o}_{i1}, \boldsymbol{o}_{i2}, \ldots, \boldsymbol{o}_{iT_i}\}$ are given by

$$\begin{aligned} \bar{\boldsymbol{\mu}}_m \quad = \quad & \Big(\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_{im}(t) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{A}_i \Big)^{-1} \times \\ & \Big(\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_{im}(t) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{o}_{it} - \boldsymbol{b}_i) \Big), \end{aligned} \tag{8}$$

$$\bar{\boldsymbol{\Sigma}}_m = \frac{\displaystyle\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_{im}(t) (\boldsymbol{o}_{it} - \tilde{\boldsymbol{\mu}}_{im})(\boldsymbol{o}_{it} - \tilde{\boldsymbol{\mu}}_{im})^\top}{\displaystyle\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_{im}(t)}, \tag{9}$$

where $\gamma_{im}(t)$ is the probability that the observation vector $\boldsymbol{o}_{it}$ is generated in m-th state at time t, and $\tilde{\boldsymbol{\mu}}_{im} = \boldsymbol{A}_i \bar{\boldsymbol{\mu}}_m + \boldsymbol{b}_i$ is the mean vectors of the Gaussian pdf adapted to speaker i using the updated regression matrix and the mean vector.

3. EXPERIMENTS

3.1. Experimental Conditions

We used a set of phonetically balanced sentences of ATR Japanese speech database for training HMMs. Based on phoneme labels and linguistic information included in the database, we made context dependent phoneme labels. We used 42 phonemes including silence and pause.

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 [11]. 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 average voice model was trained using 150 sentences for each speaker from 3 female and 3 male speaker's speech data. We set the weight for adjusting the number of parameters of the model in STC as $c = 0.4$. In SAT, one regression matrix was used for each speaker and was estimated only once. For comparison, we also trained the average voice models with applying STC only and SAT only, respectively. Table 1 shows the total number of distributions included in the average voice models after clustering. The entries for "NONE", "SAT", "STC", and "STC+SAT" correspond to the obtained models using the conventional technique [1], SAT only, STC only, and the proposed technique, respectively.

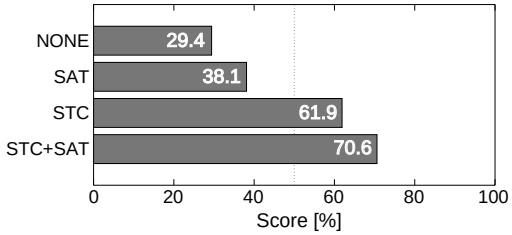


Fig. 2. Evaluation of naturalness of average voice.

We chose a female speaker FTK and a male speaker MMY from the database as the target speakers, who were not included in the training speakers of the average voice model. Based on MLLR-based speaker adaptation technique of [1], the average voice models were adapted to the target speaker using 10 sentences which were not included in the training data sentence set. In the speaker adaptation, thresholds for traversing regression class tree were set to 1000 for spectrum stream and 100 for F_0 stream, respectively. We did not adapt state duration distributions and used the same distributions as the average voice model.

3.2. Subjective Evaluations of Average Voice

We compared the naturalness of the average voice models by a paired comparison test. Subjects were 9 persons, and presented a pair of average voices synthesized from different models in random order and then asked which average voice sounded more natural. For each subject, five test sentences were chosen at random from 53 test sentences which were not contained in the training and adaptation data sentence set.

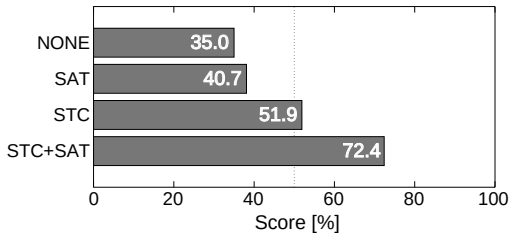
Figure 2 shows the preference scores. It can be seen from the figure that the proposed technique, namely applying both STC and SAT, provides the highest performance. In fact, we have observed that the proposed technique reduces unnaturalness of the average voice speech especially in prosodic features.

3.3. Subjective Evaluations of Adapted Voice

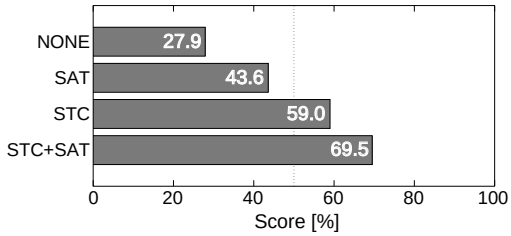
We evaluated naturalness of the synthesized speech generated from the models adapted to the given target speaker. Subjects were 7 persons. Other experimental conditions were same as the evaluation test described in the previous section.

Figure 3 shows the preference scores. In the figure, (a) is the result for a male target speaker MMY, and (b) is for a female target speaker FTK. It can be seen that similar results as the average voice were obtained for the synthesized speech from the adapted models. This means that the quality of the average voice crucially affects the quality of synthesized speech from adapted model. Moreover, the proposed technique improves the performance compared with the conventional method.

We then conducted a Comparison Category Rating (CCR) test to evaluate voice characteristics of synthesized speech from adapted models. Seven persons listened to 8 sentences of synthesized speech chosen randomly from 53 test sentences and rated their voice characteristics and prosodic features comparing to those of the reference speech. The reference speech was synthesized by a mel-

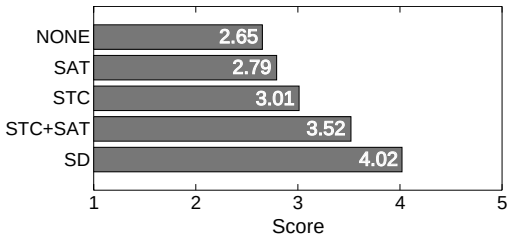


(a) male speaker : MMY

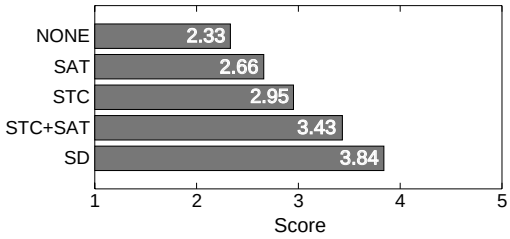


(b) female speaker : FTK

Fig. 3. Evaluation of naturalness of adapted voice.



(a) male speaker : MMY



(b) female speaker : FTK

Fig. 4. Evaluation of speaker characteristics of adapted voice.

cepstral vocoder. The rating is a 5-point scale, that is, 5 for very similar, 4 for similar, 3 for slightly similar, 2 for dissimilar, and 1 for very dissimilar. For comparison, we also evaluated synthesized speech with using speaker dependent units of the target speakers FTK and MMY. Each speaker dependent model was trained using 450 sentences uttered by the target speaker. The total numbers of distributions of the speaker dependent model for MMY were 833, 1410, and 1399 for spectrum, F_0, and state duration, respectively, and those for FTK were 891, 2057, and 1222, respectively.

Figure 4 shows the result of the CCR test. In the figure, (a) is the result for the target speaker MMY and (b) is for FTK. The score for "SD" corresponds to the result for synthesized speech using the speaker dependent model of the target speaker. This result confirms again that the proposed technique provides higher performance than the conventional techniques. Moreover, it is noted that the score for the proposed technique is close to that for the speaker dependent model.

4. CONCLUSION

We have described a new training method of average voice model for speech synthesis using speaker adaptation. The proposed training method is based on STC and SAT to reduce influence of speaker dependence and improve the quality of the synthetic speech. From the results of subjective tests, we have shown that voice characteristics of synthetic speech generated from the adapted model using the proposed method is closer to the target speaker than the conventional method. Future work will focus on application of the proposed technique to speaking style.

5. REFERENCES

[1] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP 2001*, May 2001, pp. 805–808.

[2] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *Proc. EUROSPEECH 2001*, Sept. 2001, pp. 345–348.

[3] C.J.Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[4] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP-96*, May 1996, pp. 389–392.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH-99*, Sept. 1999, pp. 2374–2350.

[6] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice model in hmm-based speech synthesis," in *Proc. ICSLP 2002*, Sept. 2002, pp. 133–136.

[7] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP-99*, Mar. 1999, pp. 229–232.

[9] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Mar. 1994, pp. 307–312.

[10] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

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