Model Adaptation and Adaptive Training using ESAT Algorithm for HMM-based Speech Synthesis

Juri Isogai, Junichi Yamagishi, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan {juri.isogai, junichi.yamagishi, takao.kobayashi}@ip.titech.ac.jp

Abstract

In speaker adaptation for HMM-based speech synthesis, model adaptation and adaptive training techniques play key roles. For reducing dependency on an initial model and adapting the model to wide-ranging target speakers, we propose speaker adaptation and adaptive training algorithms based on ESAT algorithm for HMM-based speech synthesis. The ESAT algorithm estimates contributing rate of several given initial models and combines them depending on likelihood of adaptation data for the target speaker. In this study, we incorporate the ESAT algorithm into a framework of hidden semi-Markov model (HSMM) to adapt both state output and duration distributions and convert both voice characteristics and prosodic features. From the results of subjective tests, we show that the ESAT algorithm lessen the dependence of synthetic speech quality on the initial model and has the potential ability for a wider range of the target speakers.

1. Introduction

Voice conversion can enhance the value of speech synthesis and hence many approaches [1]-[4] have been introduced. We have proposed an HMM-based voice conversion approach using speaker adaptation techniques [5][6]. In this system, first, spectrum, fundamental frequency, and duration of several training speakers are modeled simultaneously in a framework of HMM and “average voice model”, which has average speaker characteristics of the training speakers, is constructed by using the adaptive training which conducts speaker normalization [7][8]. Then, the average voice model is adapted to a new target speaker with a small amount of speech data of the target speaker by using maximum likelihood linear regression (MLLR) [6][9]. After the speaker adaptation, speech is synthesized in the same way as speaker-dependent speech synthesis method using a parameter generation algorithm proposed in [10].

This system enables us to adapt not only spectral features but also prosodic features including fundamental frequency and phoneme duration. However, the quality of synthesized speech of the adapted model depends on the given initial average voice model crucially. Hence, we need to consider how to choose the average voice model appropriate to the target speaker. One of the reasonable strategies is to use gender-specific average voice model. If we are able to know the gender of the target speaker as prior information, we can use the gender-specific average voice model as an initial model. However, it is not always suitable for all of speakers to use the gender-specific average voice model because there would be male speakers whose voice and prosodic characteristics are closer to those of female average voice model, and vice versa. More reasonable strategy is to use several typical average voice models at the same time and automatically blend them depending on speaker characteristics of the target speaker.

In this study, to actualize the strategy, we incorporate ESAT algorithm [11] into our HMM-based speech synthesis system. The speaker adaptation based on ESAT algorithm creates the model of the target speaker by estimating the contributing rate of each average voice model and combines average voice models depending on likelihood of the adaptation data for the target speaker. We show that this ESAT algorithm provides reduction of the dependence on the initial model and wide-range speaker adaptation.

2. Model Adaptation based on ESAT

In speaker adaptation for the speech synthesis, we need to perform the adaptation of not only state output distributions but also duration distributions to convert both voice characteristics and prosodic features. However, it is not straightforward to adapt phoneme or segmental duration, which is one of prosodic features, in rigorous way because the original HMM does not have explicit duration distributions. Therefore, we apply ESAT algorithm [11] to a framework of hidden semi-Markov model (HSMM) [12] which is an HMM with explicit state duration probabilities. Here we describe the HSMM-based ESAT algorithm.

The state output and duration distributions of an initial model are Gaussian distributions characterized by mean vector \( \mu_c \) and diagonal covariance matrix \( \Sigma_c \), and mean \( m_{ci} \) and variance \( \sigma^2_{ci} \), respectively. Let \( C \) be the total number of the initial model, and \( \tilde{\mu}^{(f)} \) and \( \tilde{m}^{(f)} \) be adapted mean vectors of output and duration distributions for the target speaker \( f \), respectively. In the adaptation using HSMMM-based ESAT algorithm, the adapted mean vectors are estimated by linearly transforming mean vector of output and duration distributions of each initial model and summing them

\[
\tilde{\mu}^{(f)} = \sum_{c=1}^C A_c^{(f)} \mu_c + b^{(f)} = W^{(f)} \xi,
\]

\[
\tilde{m}^{(f)} = \sum_{c=1}^C \chi_c^{(f)} m_{ci} + \nu^{(f)} = X^{(f)} \phi,
\]

where \( i \) indicates the index number of the state, \( W^{(f)} = [A_1^{(f)}, \ldots, A_C^{(f)}, b^{(f)}] \) and \( X^{(f)} = [\chi_1^{(f)}, \ldots, \chi_C^{(f)}, \nu^{(f)}] \) are \( n \times (Cn + 1) \) and \( 1 \times (C + 1) \) transformation matrices for state output and duration distributions, respectively, and \( \xi = [\tilde{\mu}_1^{(f)}, \ldots, \tilde{\mu}_C^{(f)}]^T \) and \( \phi = [m_{11}, \ldots, m_{1n}, \nu_1] \) are the \( (Cn + 1) \)-dimensional and \( C + 1 \)-dimensional vectors, re-
respectively. Note that $A_{i}^{(f)}$ is $n \times n$ matrix which transforms the output distribution of the initial model $c$ and $\chi_{c}^{(f)}$ is for the duration distribution of the initial model $c$.

The optimum transformation matrices are estimated so as to maximize the likelihood of the adaptation data for the target speaker $f$. The re-estimation formulas are given by

$$\bar{\mathbf{w}}_{i}^{(f)\top} = G_{i}^{-1} \mathbf{y}_{i},$$

$$\bar{\mathbf{X}}^{(f)} = \left( \sum_{t,r,d} \frac{\gamma_{i}^{d}(r) d}{\sigma_{i}^{2}} \phi_{i}^{T} \right)^{-1} \left( \sum_{t,r,d} \frac{\gamma_{i}^{d}(r) d}{\sigma_{i}^{2}} \phi_{i} \phi_{i}^{T} \right)^{-1} \left( \sum_{t,r,d} \frac{\gamma_{i}^{d}(r) d}{\sigma_{i}^{2}} \phi_{i} \right)$$

where $T_{f}$ is length of adaptation data for the speaker $f$, $\gamma_{i}^{d}(r)$ is state occupancy probability of HSMM [8], $(Cn+1) \times (Cn+1)$ matrix $G_{i}$ and $n \times (Cn+1)$ matrix $Y$ is given by

$$G_{i} = \sum_{t,r,d} \gamma_{i}^{d}(r) \Sigma_{(r)}^{-1} \xi_{i,t},$$

$$Y = \sum_{t,r,d} \gamma_{i}^{d}(r) \Sigma_{(r)}^{-1} \sum_{s=t-d+1}^{t} \alpha_{i,s} \xi_{i,t}^{s},$$

where $\mathbf{w}_{i}^{(f)}$ and $\mathbf{y}_{i}$ are the $l$-th row vectors of $\mathbf{W}_{i}^{(f)}$ and $\mathbf{Y}$ respectively, $\Sigma_{(r)}$ is the $l$-th diagonal element of $\Sigma_{r}$. Note that the transformation matrices are tied across $R$ distributions.

3. Adaptive Training based on ESAT

We can also apply the ESAT algorithm to adaptive training [9][11] which conducts normalization of speaker differences and acoustic variability in both output and state duration distributions of the initial models.

Let $F$ be the total number of training speakers and $O_{f}^{(i)} = \{o_{f,i}, \ldots, o_{f,T} \}$ be the training data of length $T_{f}$ for training speaker $f$. The re-estimation formulas for the mean vectors are given by

$$\bar{\xi}_{i} = \frac{1}{\sum_{f,t} \gamma_{f}^{d}(i) \mathbf{W}_{i}^{(f)\top} \Sigma_{(f)}^{-1} \mathbf{W}_{i}^{(f)}}$$

$$= \frac{\sum_{f,t} \gamma_{f}^{d}(i) \mathbf{W}_{i}^{(f)\top} \Sigma_{(f)}^{-1} \sum_{s=t-d+1}^{t} (\alpha_{i,s} - \bar{\mathbf{b}}^{(f)})}{\sum_{f,t} \gamma_{f}^{d}(i) \mathbf{W}_{i}^{(f)\top} \Sigma_{(f)}^{-1} \sum_{s=t-d+1}^{t} \alpha_{i,s}}$$

$$\bar{\psi}_{i} = \frac{1}{\sum_{f,t} \gamma_{f}^{d}(i) \mathbf{X}_{i}^{(f)\top} \mathbf{X}_{i}^{(f)}}$$

$$= \frac{\sum_{f,t} \gamma_{f}^{d}(i) (\mathbf{X}_{i}^{(f)} - \bar{\mathbf{X}}_{i}) \mathbf{X}_{i}^{(f)\top}}{\sum_{f,t} \gamma_{f}^{d}(i) (\mathbf{X}_{i}^{(f)} - \bar{\mathbf{X}}_{i}) \mathbf{X}_{i}^{(f)\top}}$$

where $\bar{\mathbf{W}}_{i}^{(f)} = \left[ \bar{\mathbf{X}}_{1}^{(f)}, \ldots, \bar{\mathbf{X}}_{C}^{(f)}, \bar{\mathbf{X}}_{1}^{(f)} \right] \mathbf{X}_{i}^{(f)} = \left[ \bar{\mathbf{X}}_{1}^{(f)}, \ldots, \bar{\mathbf{X}}_{C}^{(f)} \right]$, and $\bar{\mathbf{X}}_{i} = \left[ \bar{\mathbf{X}}_{1}, \ldots, \bar{\mathbf{X}}_{C} \right]$. The re-estimation formulas for the covariance matrices are identical to HSMM-based SAT [8]. This algorithm also becomes the HSMM-based MLLR/SAT [6][8] in $C = 1$.

4. HSMM-Based Speech Synthesis System using ESAT Algorithm

In this study, we use an HSMM-based speech synthesis system. The basic structure is similar to the HMM-based speech synthesis system of [7].

In the training stage, context dependent phoneme HSMMs are trained using multi-speaker speech database. Spectrum, $F_{0}$, and duration are modeled by multi-stream HSMMs in which output distributions for spectral and $F_{0}$ parts are modeled using continuous probability distribution and multi-space probability distribution [13], respectively. To model variations of spectrum, $F_{0}$, and duration, we take several phonetic and linguistic contextual factors into account. Then, shared decision tree based clustering technique [7] is separately applied to the spectral, $F_{0}$, and duration parts of the context dependent HSMMs. Moreover, we apply re-estimation process using the ESAT algorithm described in Sect.3 to the clustered and tied context dependent HSMMs. We trained two gender-dependent HSMMs as the initial models of the adaptation.

In the adaptation stage, the gender-dependent initial models are adapted to a target speaker using a small amount of speech data uttered by the target speaker. We use the ESAT algorithm described in Sect.2 to adapt spectrum, $F_{0}$, and state duration at the same time.

In the synthesis stage, texts are transformed into a context dependent label sequence. In accordance with the label sequence, a sentence HSMM is constructed by concatenating context dependent HSMMs. From the sentence HSMM, spectral and $F_{0}$ parameter sequences are obtained based on ML criterion. Finally, by using MLSA filter, speech is synthesized from the generated mel-cepstral and $F_{0}$ parameter sequences.

5. Experiments

5.1. Experimental Conditions

To evaluate the effectiveness of the proposed techniques, we conducted some subjective evaluation tests for synthetic speech using speaker adaptation. The proposed techniques were compared to the conventional HSMM-based TTS system using HSMM-based MLLR and SAT [8].

We used a set of phonetically balanced sentences of ATR Japanese speech database (Set B) for training data of HSMMs. 42 phonemes including silence and pause were used. Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. 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 HSMMs. We chose two male speakers MHT and MTK, and a female speaker FTK from the database as the target speakers, who were not included in the training speakers of the average voice model. The average voice models were adapted to the target speaker using 50 sentences which were not included in the training data sentence set. We adopt two gender dependent models as initial models. The each gender dependent initial model was trained using 1350 sentences, 450 sentences for each of three male or female speakers. For comparison, we also trained a gender independent model using 2700 sentences, 450 sentences for each of the six speakers. Furthermore, we also trained a speaker dependent model using 450 sentences for the target speaker.

5.2. Comparison of formation of transformation matrices

We first compared the synthesized speech samples generated from two transformation matrices in different forms; full matrix and block diagonal matrix. Here, the initial models ($C = 2$) were adapted to a male speaker MHT by using ESAT. Figure 1 shows the mel-cepstral distance between spectra generated from
For the distance calculation, 53 test sentences were used for evaluation, which were included in neither training nor adaptation data, state duration was adjusted after Viterbi alignment with the target speaker’s real utterance. Note that silence and pause regions were eliminated in the distance calculation.

In this figure, “FULL” represents the result for the full transformation matrix and “BLOCK” represents that for the block diagonal transformation matrix. From the figure, we can see that the block diagonal matrix outperforms the full transformation matrix when the amount of adaptation data is small. From this result, we decided to use the block diagonal matrix in the following subjective evaluations.

5.3. Subjective Evaluation for Naturalness of Synthesized Speech

We evaluated the naturalness of the synthesized speech generated from the models adapted to the target speakers MHT, MTK, and FTK by a paired comparison test. Subjects were 9 persons, and presented a pair of speech samples generated using different techniques in random order and then asked which samples had better naturalness. For each subject, three test sentences were chosen at random from 53 test sentences which were contained in neither training nor adaptation data sentence set.

Figure 2 shows the preference scores. A confidence interval of 95% is also shown in the figure. In the figure, (a) is the result for the target speaker MHT, (b) is that for the target speaker MTK, and (c) is that for the target speaker FTK. In the figure, “GD_M” and “GD_F” represent the results for the male- and female-dependent model using SAT [8] and MLLR adaptation [6], respectively, “SAT+ESAT” represents the results for the models using SAT and ESAT-based adaptation described in Sect.2, “ESAT+ESAT” represents the results for the models using the ESAT algorithm described in Sect.3 and the ESAT adaptation, and “GI” represents the results for a gender independent model using the SAT and MLLR. From the figure, we can see that the results of the conventional techniques depend on the given initial models. In contrast, the proposed method (ESAT+ESAT) lessen the dependence of synthetic speech quality on the initial model and provides naturalness comparable to that of the gender dependent model for each target speaker.

5.4. Subjective Evaluation for Speaker Characteristics of Synthesized Speech

We then conducted a comparison category rating (CCR) test to evaluate speaker characteristics of synthesized speech from adapted models. Six persons listened to 8 sentences of synthesized speech chosen randomly from 53 test sentences and rated their speaker characteristics comparing to those of the reference speech. The reference speech was synthesized by a mel-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 using speaker dependent models of the target speakers MHT, MTK, and FTK.

Figure 3 shows the result of the CCR test. In the figure, (a) is the result for the target speaker MHT, (b) is that for MTK, and (c) is that for FTK. The score for “SD” corresponds to the result for synthesized speech using the speaker dependent model of the target speakers MHT, MTK, and 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 ESAT algorithm lessen the dependence of synthetic speech quality on the initial model and has the potential ability of improving adaptation performance for a wider range of the target speakers.
6. Conclusions

This paper have described a speaker adaptation technique based on ESAT algorithm for HSM-M-based speech synthesis. Moreover, we have derived an adaptive training algorithm based on ESAT. From the results of subjective tests, we have shown that the ESAT algorithm lessen the dependence of synthetic speech quality on the initial model and has the potential ability for a wider range of the target speakers. Future work will focus on the evaluation of the proposed method using wide-ranging target speakers.

7. Acknowledgments

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (B) 15300055 and JSPS Research Fellowships for Young Scientists 164633.

8. References


