MODEL ADAPTATION APPROACH TO SPEECH SYNTHESIS WITH DIVERSE VOICES AND STYLES

Junichi Yamagishi†,††, *Takao Kobayashi*†, *Makoto Tachibana*†, *Katsumi Ogata*†, *Yuji Nakano*†

†Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan
††The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9LW United Kingdom

ABSTRACT

In human computer interaction and dialogue systems, it is often desirable for text-to-speech synthesis to be able to generate natural sounding speech with an arbitrary speaker's voice and with varying speaking styles and/or emotional expressions. We have developed an average-voice-based speech synthesis method using statistical average voice models and model adaptation techniques for this purpose. In this paper, we describe an overview of the speech synthesis system and show the current performance with several experimental results.

Index Terms— HMM, speech synthesis, average voice, speaker adaptation, voice conversion

1. INTRODUCTION

State-of-the-art concatenative speech synthesis methods using large-scale speech corpora give us high quality synthetic speech. However, as is clear from the principle of the methods, they face the problem that a large-scale corpus is always required for each speaker or speaking style. Hence, in order to develop a humanlike speech synthesizer which can control many kinds of speaking style using the concatenative speech synthesis method, we would be required to prepare many corpora corresponding to the separate speaking styles. Furthermore, if we need to construct a speech synthesizer which can deal with many speakers' voices, we would also be required to prepare immense corpora corresponding to all combinations of the speakers and speaking styles. To make speech synthesis with diverse voices and styles realistically feasible, we should make the amount of the speech data required for a new speaker or new speaking style as small as possible. Moreover, it is preferable that the quality of synthetic speech is comparable to that of a speaker-dependent system built using a large amount of speech data.

For the purpose of constructing a novel speech synthesis system which can achieve diverse voices and styles, we have been developing a statistical speech synthesis method using average voice models created by hidden semi-Markov model (HSMM), model adaptation techniques, and several relevant techniques (e.g. [1][2]). By using this speech synthesis framework, synthetic speech of the target speaker can be obtained robustly even when the amount of speech data available from the target speaker are very small. The speech synthesis method will be referred to as "average-voice-based speech synthesis" in the following section.

In this paper, we give an overview of the up-to-date system based on our recent research achievements and show the current performance and issues based on several experimental results.

2. AVERAGE-VOICE-BASED SPEECH SYNTHESIS

The average-voice-based speech synthesis framework consists of a speech analysis part, acoustic modeling part, model training and speaker normalization part for the average voice model, speaker adaptation part, and speech synthesis part.

2.1. Speech Analysis

We use the mel-cepstrum and logF0 as acoustic features. The mel-cepstral coefficients are obtained by mel-cepstral analysis [3], and F0 values are estimated using an IFAS (instantaneous frequency amplitude spectrum) based method [4]. Then dynamic and acceleration features are calculated as the first and second order regression coefficients from the current frame and their adjacent frames.

2.2. Acoustic Modeling

To simultaneously model the estimated acoustic features and duration in a unified modelling framework using a consistent criterion, we utilize context-dependent multi-stream left-to-right MSD (multi-space distribution) [5] HSMMs [6]. The phonetic and linguistic contexts which we used in the following experiments contain phoneme, mora, part-of-speech, accentual information, breath group, and sentence information. The multi-stream is used for simultaneously modelling mel-cepstrum and logF0. Then the MSD enables us to treat F0 observation, which is a mixture observation of 1-dimensional real number for voiced region and symbol string for unvoiced region [5], as a probability framework. Finally the HSMMs enable us to explicitly estimate duration distributions in addition to mel-cepstrum and logF0 [6] instead of the transition probability in the framework of the original HMM. Note that the phoneme boundary labels are used only for obtaining the initial model parameters of the average voice model and we do not require the phoneme boundary labels in the adaptation or synthesis stage.

2.3. Model Training and Speaker Normalization

Using the above HSMMs, we train an average voice model as the initial model of the adaptation from training data which consists of several speakers' speech. We then adapt the model to that of a target speaker by using a small amount of speech data uttered by the target speaker.

The training data of the average voice model includes a lot of speaker- and/or gender-dependent characteristics and they crucially affect the adapted models and the quality of synthetic speech generated from them. For constructing an appropriate average voice model, we should deal with the negative effects caused by the speaker- and/or gender-dependent characteristics when we estimate or cluster the model parameters for the acoustic features.

Therefore, we utilize the SAT (speaker-adaptive training) algorithm for normalizing the influence of speaker differences in the parameter estimation for the average voice model [7]. In the SAT algorithm, the speaker difference is assumed to be expressed as piecewise linear regression functions of the average voice model, and the model parameters for the average voice model are blindly estimated so that the model parameters and regression matrices maximize the likelihood for the training data.

We then conduct a speaker normalization technique called the STC (shared-tree-based clustering) algorithm [7] in the tree-based

clustering of the model parameters for the average voice model. Using this technique, every node of the decision tree always has statistics from all training speakers. As a result, we can construct decision trees common to all training speakers and each parameter of the node always reflects the statistics of all speakers.

2.4. Speaker Adaptation

In the speaker adaptation stage, we adapt the average voice model to that of the target speaker by using a small amount of speech data. In speaker adaptation for speech synthesis, it would be desirable to simultaneously convert mel-cepstrum, F0, and duration. Therefore, we utilize speaker adaptation techniques for the multi-stream MSD-HSMM [8]. Many speaker adaptation algorithms, including SMAP, MLLR, CMLLR, SMAPLR, and CSMAPLR [9] can be utilized in the HSMM [1]. Such speaker adaptation algorithms utilize piecewise linear regression functions. For automatic determination of tying topology of the regression matrices for the piecewise linear regression, we utilize shared-decision-trees [8] because the decision trees have phonetic and linguistic contextual questions related to the suprasegmental features by which prosodic features, especially F0, are characterized.

The above speaker adaptation algorithms have a rough assumption that the target speaker model would be expressed by the piecewise linear regression of the average voice model. Therefore, we additionally adopt MAP (Maximum A Posteriori) modification [10] to upgrade the adapted model parameters which have a relatively large amount of speech data from the target speaker.

2.5. Speech Synthesis

In the synthesis stage, an arbitrarily given text is first transformed into a sequence of context-dependent phoneme labels. Based on the label sequence, a sentence HSMM is constructed by concatenating context-dependent HSMMs. From the sentence HSMM, mel-cepstrum and logF0 sequences are obtained using the parameter generation algorithm [11], in which phoneme durations are determined using state duration distributions. Finally, by using an MLSA filter [3], speech is synthesized from the generated mel-cepstrum and logF0 sequences.

3. EXPERIMENTS

3.1. Experimental conditions

We conducted several objective and subjective evaluation tests. We used the ATR Japanese speech database (Set B), which contains a set of 503 phonetically balanced sentences uttered by 6 male speakers (MHO, MHT, MMY, MSH, MTK, MYI) and 4 female speakers (FKN, FKS, FTK, FYM), and a speech database which contains the same sentences as the ATR Japanese speech database uttered by a male speaker (MMI) and a female speaker (FTY). In the modeling of the synthesis units, we used 42 phonemes, including silence and pause, and took the phonetic and linguistic contexts [7] into account. The contexts are manually labeled based on real speech from each speaker.

Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Blackman window with a 5-ms shift. The feature vectors consisted of 25 mel-cepstral coefficients (including the zeroth coefficient), logF0, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs without skip paths. Gender-dependent average voice models were trained from multiple training speakers. The number of training sentences for each training speaker was 453 sentences, and details of the training and target speakers are written in each subsection. In the training stage of the average voice models, the number of leaf nodes of the shared decision trees was determined using the minimum description length (MDL) criterion. In the speaker adaptation and speaker-adaptive training, multiple regression matrices were estimated based on the shared-decision-trees. The threshold that specifies an expected number of speech samples used for each regression matrix was determined from preliminary objective experimental results. The transformation matrices were diagonal triblock corresponding to the static, dynamic, and acceleration coefficients. In the following assessments, fifty test sentences which were not included in either the training or the adaptation data were used for the evaluation.

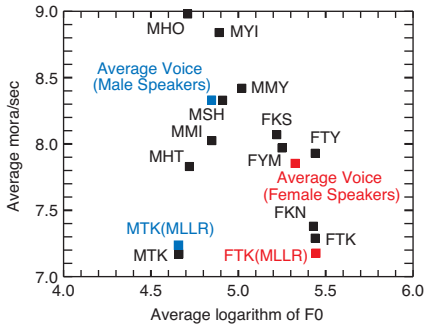

Fig. 1. Distribution of average logF0 and mora/sec for each speaker, average voice generated from the average voice mode and synthetic speech generated from the adapted model using 10 sentences.

3.2. Evaluation of average logF0 and duration

First, we evaluated the average logF0 and mora/sec of the synthetic speech generated from the adapted model. We chose a male speaker MTK and a female speaker FTK as target speakers of the speaker adaptation and used 5 male speakers (MHO, MHT, MMY, MSH, MYI) and 4 female speakers (FKN, FKS, FYM, FTY) as training speakers for the average voice model. The adaptation method used in this experiment was MLLR. Figure 1 shows the average logF0 and mora/sec of real speech of each speaker, the average voice (synthetic speech generated from the average voice model) and synthetic speech generated from the adapted model using 10 sentences. From the figure, we can see that the average logF0 and mora/sec of the synthetic speech generated from the adapted model are close to those of the target speakers' speech.

3.3. Objective evaluation of synthetic speech

Next we evaluated average mel-cepstral distance and RMSE of logF0 and vowel duration between real and synthetic speech of the target speakers as the objective measures. Training speakers for the average voice model and target speakers are the same as Section 3.2. The adaptation method was MLLR, and the number of the adaptation sentences was from 3 sentences to 450 sentences. For the calculation of the average mel-cepstral distance and the RMSE of logF0, the state duration of each HSMM model was adjusted after Viterbi force-alignment with the target speakers' real utterance. In the mel-cepstral distance calculation, silence and pause regions were eliminated. In the RMSE calculation of logF0, only regions where both the generated and the real F0 were voiced were used since F0 is not observed in the unvoiced region. In the RMSE calculation of vowel duration, the manually labeled duration of real speech was used as the target vowel duration.

Figure 2 shows the evaluation results using these objective measures between real and synthetic speech of the target speakers. In this figure, zero sentence means those of the male- or female-speaker average voice models. From these figures, we can see that all features of synthetic speech generated from the adapted model become closer to the target speakers' features than those of average voice model by using a small amount of speech data. Comparing the adaptation of logF0 and mel-cepstrum, we see that the improvement of RMSE of logF0 converges when just a few adaptation sentences are used,

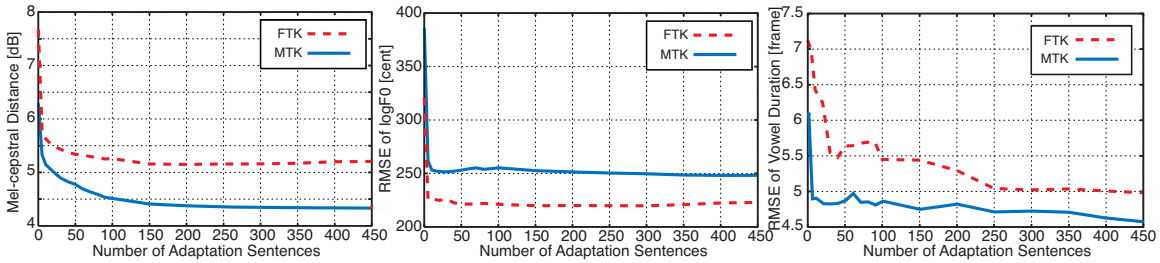

Fig. 2. Left: Average mel-cepstral distance [dB], Center: RMSE of logF0 [cent], Right: RMSE of vowel duration [frame].

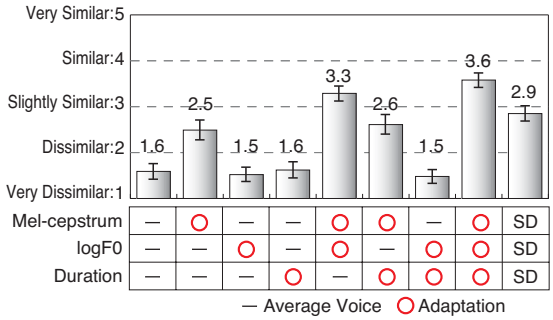

Fig. 3. Subjective evaluation of adaptation effects of each feature.

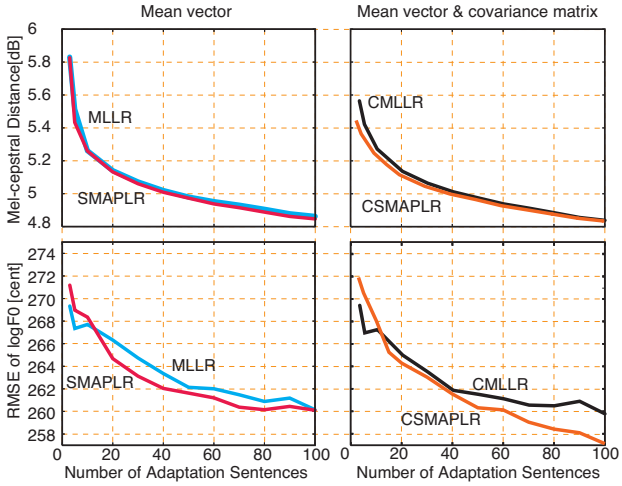

Fig. 4. Objective evaluation of speaker adaptation algorithms.

whereas about 50 to 150 sentences are needed for the convergence of mel-cepstral distance. This is due to the different numbers of parameters for the features. Meanwhile, in the adaptation of the duration, the required number of adaptation sentences varies with the target speaker.

3.4. Subjective evaluation of synthetic speech

Next, we conducted a comparison category rating (CCR) test and assessed the effectiveness of the adaptation of each feature. We compared the synthesized speech generated from eight models with or without the adaptation of mel-cepstrum, logF0, and/or duration. Training speakers for the average voice model and target speakers are the same as Section 3.2. The adaptation method was MLLR, and the adaptation data comprised 100 utterances. For reference, we also compared synthesized speech generated from a speaker-dependent (SD) model [6] using 453 sentences of the target speaker. Eight subjects were first presented with the reference speech sample and then with synthesized speech samples generated from the adapted models in random order. The subjects were asked to rate their voice characteristics and prosodic features compared with those of the reference speech. The reference speech was synthesized with a mel-cepstral vocoder. The rating was done on 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 each subject, five test sentences were randomly chosen.

Figure 3 shows the average values with 95 % confidence interval of the CCR tests. The values indicate that it is preferable to simultaneously adapt all features to reproduce the speaker characteristics of the target speaker. It is interesting to note that the synthesized speech generated from the model using the simultaneous adaptation of all features using 100 target-speaker's sentences is rated as having more similar speaker characteristics to the target speaker compared with the speaker-dependent models using 453 sentences of the target speaker. We will describe the reason in Section 3.6.

3.5. Objective evaluation of adaptation method

We then compared several adaptation methods by using the objective measures described in Section 3.3. We chose three male speakers MHT, MTK, MMI and a female speaker FTK as target speakers of the speaker adaptation and used 4 male speakers (MHO, MMY, MSH, MYI) and 4 female speakers (FKN, FKS, FYM, FTY) as training speakers for the average voice model. All other conditions are the same as in Section 3.3. To compare the transformation function and criterion, we chose four adaptation methods – MLLR, CMLLR, SMAPLR, and CSMAPLR. The transformation function of MLLR and SMAPLR is composed of linear functions of the mean vectors of Gaussian pdfs, and that of CMLLR and CSMAPLR is composed of linear functions of the mean vectors and covariance matrices of Gaussian pdfs. In MLLR and CMLLR, the ML criterion is used, whereas the Structural MAP criterion [9] is used in SMAPLR and CSMAPLR. We used 100 adaptation sentences for each method. Figure 4 shows the evaluation results using the objective measures between real and synthetic speech of the target speakers. From this figure, it can be seen that the CSMAPLR algorithm using the functions of the mean vectors and covariance matrices together with the SMAP criterion generally brings about an improvement in the RMSE of logF0, whereas it slightly decreases the average mel-cepstral distance.

Next we applied the MAP modification [10] to the model created by CSMAPLR adaptation and compared the speech before-and-after. Figure 5 shows the evaluation results using the objective measures between real and synthetic speech with and without the MAP modification of the target speakers. The result confirms that as the number of available adaptation sentences increases, the improvement due to MAP modification also increases. The MAP modification seems to have a beneficial effect on the improvements of RMSE of logF0.

3.6. Comparison of speaker-dependent and adapted models

If the decision trees of the adapted model with MAP modification and the speaker dependent model are identical, the objective mea-

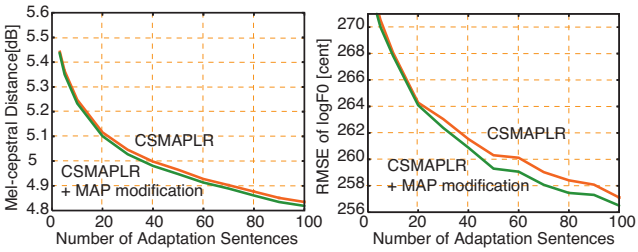

Fig. 5. Objective evaluation of MAP modification.

sures of the adapted model would asymptotically converge to that of the speaker dependent model using the ML criterion. However, since we construct the decision tree from the training data of the average voice model as described in Section 2.3, the convergence is not guaranteed theoretically. In addition, because a huge amount of speech data would be required to obtain the optimal speaker-dependent model and decision trees, it is likely that the adapted model using the average voice model (which is trained from a large amount of speech data from many speakers) can generate synthetic speech more stably and robustly using a realistic and practical amount of speech data [10].

Therefore, we compared the speaker-adapted model with MAP modification and the speaker-dependent model by using the objective measures described in Section 3.3. We chose a male- and a female- target speaker from a set of 6 male speakers (MHO, MHT, MMY, MSH, MTK, MYI) and 5 female speakers (FKN, FKS, FYM, FTK, FTY) and used the rest of the speakers as training data. We then repeatedly conducted the evaluations to different target speakers. In this evaluation, in order to verify the effect of the number of training sentences for the average voice model, we used a gender-independent average voice model and changed the total number of training data for the average voice model from 450 sentences to 4050 sentences. We set the number of the adaptation sentences to 450 sentences to compare the speaker-dependent and adapted models. Figure 6 shows the evaluation results using the objective measures between real and synthetic speech of the target speakers. In this figure, "SD-*" and "SA-*" signify the speaker-dependent and adapted models respectively. We can see that as the number of training sentences for the average voice model increases, both the mel-cepstral distance and RMSE of logF0 generally decrease. We also see that when the average voice models are trained from a large amount of speech data, the RMSE of logF0 of the adapted models is smaller than that of the speaker-dependent model in all the target speakers, whereas significant differences do not exist in the average mel-cepstral distance. From several subjective evaluations, we also confirmed that when the average voice models are trained from a large amount of speech data, the average voice models have more model parameters than the speaker-dependent model, and synthetic speech of the adapted model outperformed that of the speaker-dependent model.

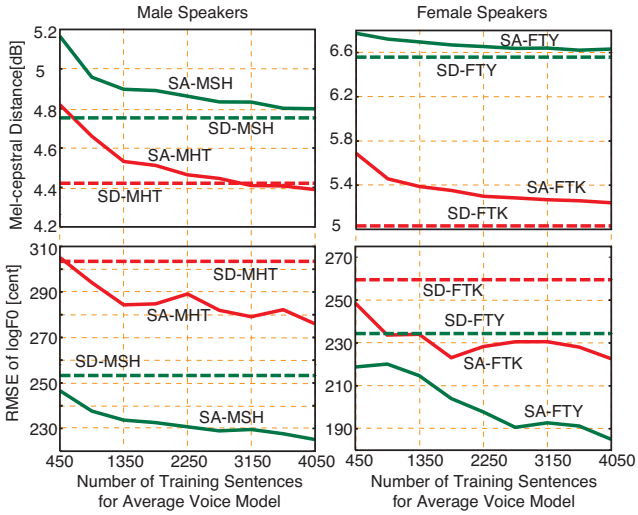

Fig. 6. Objective evaluation of speaker-dependent and adapted models.

4. CONCLUSIONS

We described and evaluated the average-voice-based speech synthesis method for generating synthetic speech with diverse speakers' voice from a small amount of speech data. Once the average voice model is constructed from a large amount of speech data, we can easily create a new speaker's synthetic speech which is comparable to or better than the speaker-dependent model. Besides speaker conversion, we can easily apply this speech synthesis method to conversion of speaking style and emotional expression [8].

5. REFERENCES

[1] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," in *Proc. ICASSP 2006*, May 2006, pp. 77–80.

[2] J. Yamagishi and T. Kobayashi, "Adaptive training for hidden semi-Markov model," in *Proc. ICASSP 2005*, Mar. 2005, pp. 365–368.

[3] 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.

[4] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust f0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Information and Systems*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.

[5] 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.

[6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. ICSLP 2004*, Oct. 2004, pp. 1393–1396.

[7] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.

[8] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Information and Systems*, vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[9] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. ICSLP 2006*, Sept. 2006, pp. 2286–2289.

[10] K. Ogata, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Acoustic model training based on linear transformation and map modification for HSMM-based speech synthesis," in *Proc. ICSLP 2006*, Sept. 2006, pp. 1328–1331.

[11] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP-95*, May 1995, pp. 660–663.