HSMM-BASED MODEL ADAPTATION ALGORITHMS FOR AVERAGE-VOICE-BASED SPEECH SYNTHESIS

Junichi Yamagishi, Katsumi Ogata, Yuji Nakano, Juri Isogai, Takao Kobayashi

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

ABSTRACT

In HMM-based speech synthesis, we have to choose the modeling strategy for speech synthesis units depending on the amount of available speech data to generate synthetic speech of better quality. In general, speaker-dependent modeling is an ideal choice for a large speech data, whereas speaker adaptation with average voice model becomes promising when available speech data of a target speaker is limited. This paper describes several speaker adaptation algorithms and MAP modification to develop consistent method for synthesizing speech in a unified way for arbitrary amount of the speech data. We incorporate these adaptation algorithms into our HSMM-based speech synthesis system and show its effectiveness from results of several evaluation tests.

1. INTRODUCTION

In HMM-based speech synthesis, it is necessary to choose the modeling strategy for speech synthesis unit depending on the amount of available speech data to generate synthetic speech of better quality. In general, speaker-dependent modeling is an ideal choice for a large speech data of a target speaker, whereas speaker adaptation with average voice model [1] becomes promising when available speech data of the target speaker is limited. In this method, spectrum, fundamental frequency (F0), and duration of several training speakers are modeled simultaneously in a framework of HMM, and average voice model, which models average voice and prosodic characteristics of the training speakers, is trained by using adaptive training for the speaker normalization [1][2]. Then, using a speaker adaptation algorithm such as MLLR adaptation, the average voice model is adapted to a new target speaker based on speech data uttered by the target speaker. After the speaker adaptation, speech is synthesized in the same manner as speaker-dependent speech synthesis method [3][4]. The average voice model can utilize a large variety of contextual information included in the several speakers' speech corpus as a priori information for the speaker adaptation and provide robust basis useful for synthesizing speech of the new target speaker. As a result, stable synthetic speech can be obtained even if speech samples available for the target speaker are very small.

In this study, we explore and compare several speaker adaptation algorithms to transform the average voice model into the target speaker's model when the adaptation data for the target speaker is limited. Furthermore, we adopt "ex-post" MAP (Maximum A Posteriori) estimation to upgrade the estimation for the distributions having sufficient amount of speech samples. When sufficient amount of the adaptation data is available, the ex-post MAP estimation theoretically matches the ML estimation which is used for the training of the speaker dependent model. As a result, it is thought that we do not need to choose the modeling strategy depending on the amount of speech data and we would accomplish the consistent method to synthesize speech in the unified way for arbitrary amount of the speech data. We incorporate these adaptation algorithms into our speech synthesis system and show its effectiveness from results of subjective and objective evaluation tests.

2. SPEAKER ADAPTATION BASED ON HSMM

In speaker adaptation for speech synthesis, it is desirable to convert both voice characteristics and prosodic features such as F0 and phone duration. Therefore, we use a framework of hidden semi-Markov model (HSMM) [5] which is an HMM having explicit state duration distributions instead of the transition probabilities for directly modeling and controlling phone durations. An N-state left-to-right HSMM without skip path λ is specified by state output probability distribution $\{b_i(\cdot)\}_{i=1}^N$ corresponding to spectrum and F0, and state duration probability distribution $\{p_i(\cdot)\}_{i=1}^N$ corresponding to phone duration. In this study, we assume that the i-th state output and duration distributions are Gaussian distributions characterized by mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$, and mean m_i and variance σ_i^2, respectively,

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2) \tag{2}$$

where $\boldsymbol{o}$ is the observation data and d is the time staying in the state i. Using the framework of the hidden semi-Markov model, we can derive speaker adaptation to simultaneously transform state output and duration distributions. In the next sections, we briefly present the several speaker adaptation algorithms based on the HSMM-based framework.

3. SPEAKER ADAPTATION USING BIAS VECTOR

We firstly describe several simple adaptation algorithms estimating the difference (bias) between the target speaker and the average voice model. In the adaptation algorithms, mean vectors of the state output and duration distributions for the speaker are obtained by adding the bias vector to mean vector of the average voice model,

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i + \boldsymbol{\epsilon}, \boldsymbol{\Sigma}_i) \tag{3}$$

$$p_i(d) = \mathcal{N}(d; m_i + \nu, \sigma_i^2), \tag{4}$$

where $\boldsymbol{\epsilon}$ and ν are the bias vectors for state output and duration distributions, respectively. SBR (Signal Bias Removal) [6] estimates a

global bias vector for all distributions. In contrast, AMCC (Automatic Model Complexity Control) [7] estimates several bias vectors. Each bias vector is estimated for a cluster of distributions defined by tree structure of distributions. The number of bias vectors is controlled based on heuristic threshold or information criterion such as MDL. Furthermore, SMAP (Structural Maximum A Posteriori) [8] takes advantage of the tree structure and estimates bias vector for each distribution. In the SMAP adaptation, the bias vector for each node of tree structure is estimated based on maximum a posteriori criterion where bias vector estimated for parent node of the current node is used as a parameter of prior distribution. Recursively calculating the MAP estimation from the root node to leaf nodes of the tree structure of distributions, we finally obtain an individual bias vector for each distribution.

4. SPEAKER ADAPTATION USING LINEAR REGRESSION

4.1. Maximum Likelihood Linear Regression

Next, we describe several adaptation algorithms in which several linear regression functions are estimated to transform the average voice model into target speaker model. Here we pick up the following four kinds of linear regression algorithms – MLLR (Maximum Likelihood Linear Regression) [9], multiple linear regression [10], CMLLR (Constrained MLLR) [11], and SMAPLR (Structural Maximum A Posteriori Linear Regression) [12].

In MLLR adaptation, which is the most popular linear regression adaptation, mean vectors of state output and duration distributions for the speaker are obtained by linearly transforming mean vector of state output and duration distributions of the average voice model,

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}\boldsymbol{\mu}_i + \boldsymbol{\epsilon}, \boldsymbol{\Sigma}_i) \tag{5}$$

$$p_i(d) = \mathcal{N}(d; \chi m_i + \nu, \sigma_i^2) \tag{6}$$

where $\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}]$ and $\boldsymbol{X} = [\chi, \nu]$ are transformation matrices which transform average voice model into the target speaker for state output and duration distributions, respectively. Although the MLLR adaptation needs more parameters for the transformation compared to bias vector only, the MLLR adaptation theoretically includes the speaker adaptation using the bias vector and we can expect more appropriate transformation when the available adaptation data is enough for the number of the parameters.

4.2. Maximum Likelihood Multiple Linear Regression

In the MLLR adaptation, mean vectors for the target speakers are estimated by a simple linear regression using a single average voice model. We can extend the simple linear regression to multiple linear regression using several average voice models,

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \textstyle\sum_{f=1}^{F} \boldsymbol{\zeta}^{(f)} \boldsymbol{\mu}_i^{(f)} + \boldsymbol{\epsilon}, \boldsymbol{\Sigma}_i) \tag{7}$$

$$p_i(d) = \mathcal{N}(d; \textstyle\sum_{f=1}^{F} \chi^{(f)} m_i^{(f)} + \nu, \sigma_i^2) \tag{8}$$

where F is the number of average voice models and $\boldsymbol{\mu}_i^{(f)}$ and $m_i^{(f)}$ are the mean vectors of the f-th average voice model. This algorithm, called ESAT [10], automatically selects or blends several typical average voice models depending on speaker characteristics of the target speaker. As a result, the ESAT would widely expand the range of the target speaker of the speaker adaptation compared to a single average voice model. However, the ESAT adaptation needs F times as many parameters as the MLLR adaptation needs. Hence, if the adaptation data is not enough for the number of parameters, the accuracy of transformation matrices decreases.

4.3. Constrained Maximum Likelihood Linear Regression

The transformed parameters of the speaker adaptations described above are limited to the mean vectors of the average voice model. However, we should tune covariance matrices simultaneously to a new speaker if the variation is one of the important factor such as F0. In CMLLR adaptation, mean vectors and covariance matrices of state output and duration distributions for the target speaker are obtained by transforming the parameters at the same time as follows:

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}\boldsymbol{\mu}_i - \boldsymbol{\epsilon}, \boldsymbol{\zeta}\boldsymbol{\Sigma}_i\boldsymbol{\zeta}^\top) \tag{9}$$

$$p_i(d) = \mathcal{N}(d; \chi m_i - \nu, \chi\sigma_i^2\chi). \tag{10}$$

This adaptation algorithm tunes not only mean values but also the range of the variation to a new speaker. Because the range of the variation is one of the important factors for F0, this algorithm would conduct more appropriate adaptation of prosodic information.

4.4. Structural Maximum A Posteriori Linear Regression

The linear regression adaptation algorithms described above can also estimate several transformation matrices based on tree structure of the distributions. The number and the tying topology of transformation matrices suitable for the amount of the adaptation data is automatically decided based on the tree structure. Because prosodic feature is characterized by many suprasegmental features, we utilize context decision trees whose questions are related to the suprasegmental features, such as mora, accentual phrase, part of speech, breath group, and sentence information to determine the tying topology for the transformation matrices.

In SMAPLR adaptation, the concept of SMAP adaptation is applied to the estimation of the transformation matrices of the MLLR, that is, the recursive MAP-based estimation of the transformation matrices from the root node to lower nodes is conducted. As a result, we can make better use of the structural information and the suprasegmental information which the context decision trees have.

5. MAXIMUM A POSTERIORI MODIFICATION

Furthermore, we adopt "ex-post" MAP (Maximum A Posteriori) estimation [13]. In the previous speaker adaptation using linear regression, there is a rough assumption that the target speaker model would be expressed by the linear regression of the average voice model. Therefore, by applying the MAP estimation to the model transformed by the linear regression additionally, we can modify and upgrade the estimation for the distribution having sufficient amount of speech samples. When sufficient amount of the adaptation data is available, the ex-post MAP estimation theoretically matches the ML estimation which is used for the training of the speaker dependent model. As a result, it is thought that we do not need to choose the modeling strategy depending on the amount of available speech data and we would accomplish the consistent speech synthesis method for synthesizing speech in the unified way for arbitrary amount of the speech data.

6. EXPERIMENTS

6.1. Experimental Conditions

To compare and verify the effectiveness of each speaker adaptation algorithm, we conducted several objective and subjective evaluation tests for the synthetic speech using each speaker adaptation algorithm. Speech database for the following experiments contains 7

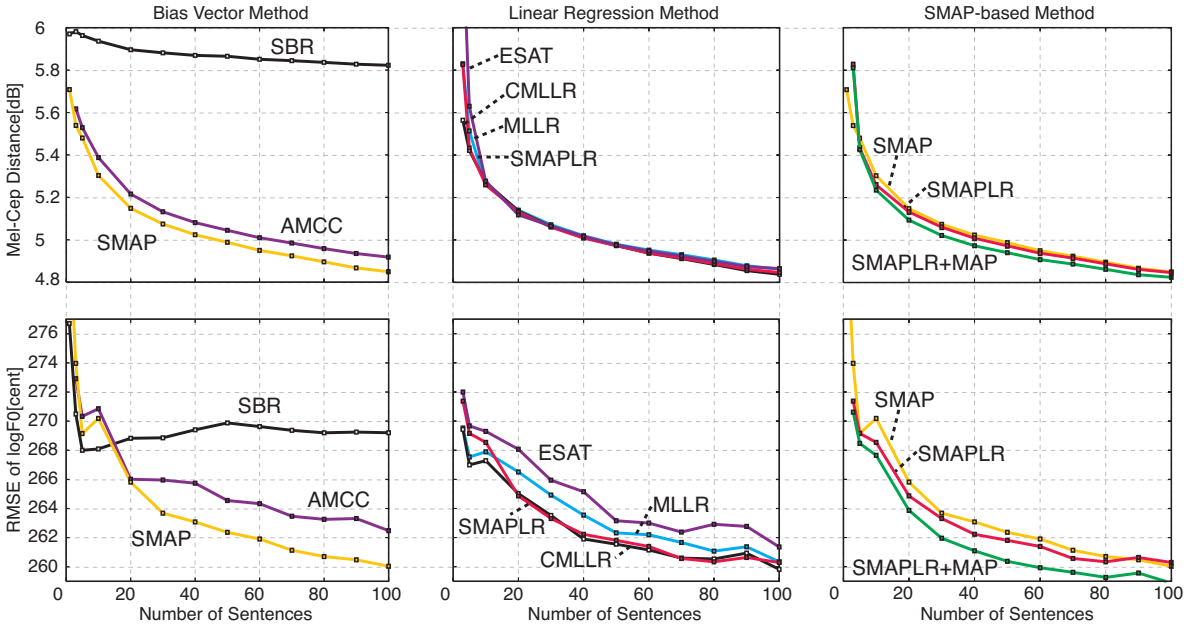

Fig. 1. Objective evaluation of speaker adaptation algorithms.

male and 5 female speakers' speech samples. Each speaker uttered a set of 503 phonetically balanced sentences taken from the ATR Japanese speech database. We chose 4 males and 4 females as training speakers for the average voice model, and used the rest of 3 males and 1 female as target speakers of the speaker adaptation. In the modeling of synthesis units, we used 42 phonemes, including silence and pause and took the phonetic and linguistic contexts [1] into account.

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 F0, and their delta and delta-delta coefficients. We used 5-state left-to-right HSMMs without skip path. The gender-dependent and independent average voice models were separately trained using 1800 and 3600 sentences, respectively, 450 sentences for each training speaker. In the training stage of the average voice models, shared-decision-tree-based context clustering algorithm and speaker adaptive training [1][2] were applied to normalize influence of speaker differences among the training speakers and train appropriate average voice models. Note that all the average voice models have the same topology and the number of distributions based on the shared-decision-trees.

We then adapted the average voice model to the target speaker using adaptation data whose sentences were included in the training sentences. In all the adaptation algorithms except SBR, multiple transformation parameters were estimated based on the shared-decision-trees constructed in the training stage of the average voice models. The tuning parameters for each adaptation algorithm, the thresholds to control the number of transformation parameters and hyper-parameters of the MAP estimation, were determined based on preliminary objective experimental results. The average voice model used as an initial model were also determined based on the preliminary objective experimental results. The gender-dependent average voice models were used for 2 male and 1 female speakers and gender-independent average voice model was used for the rest of a male speaker. ESAT adaptation used both the gender-dependent and gender-independent average voice models as the initial models.

6.2. Objective Evaluations of Speaker Adaptation Algorithms

Firstly, we calculated the target speakers' average mel-cepstral distance and root-mean-square (RMS) error of logarithmic F0 as the objective evaluations for each speaker adaptation algorithm. The number of the adaptation sentences ranged from three to a hundred. Fifty test sentences were used for evaluation, which were included in neither training nor adaptation data. For the distance calculation, state duration of each model was adjusted after Viterbi alignment with the target speaker's real utterance. Figure 1 shows the target speakers' average mel-cepstral distance between spectra generated from each model and obtained by means of analyzing target speaker's real utterance, and the RMS logarithmic F0 error between generated logarithmic F0 and that extracted from target speaker's real utterance. In the distance calculation, silence and pause regions were eliminated. And since F0 value is not observed in the unvoiced region, the RMS logarithmic F0 error was calculated in the region where both generated F0 and real F0 were voiced. From this figure, it can be seen that making better use of the structural information and the suprasegmental information which the context decision trees have based on the SMAP concept and estimating multiple transformation parameters provides better synthetic speech similar to the target speaker. And we can see that CMLLR adaptation to tune both mean and variance, and MAP modification to upgrade the estimation accuracy also have a beneficial effect on the improvements of F0 even for the case where the adaptation data is small.

Figure 2 shows the average mel-cepstral distance and the RMS logarithmic F0 error of the synthetic speech using the speaker adaptation algorithms (SMAPLR and MAP Modification) and speaker dependent (SD) algorithms [4]. The maximum number of sentences for the target speaker was 450 sentences. From this figure, we can see that the speaker adaptation algorithm significantly outperforms the speaker dependent model when the adaptation data is relatively limited, and furthermore, when relatively sufficient amount of the adaptation data is available, the error of synthetic speech using the speaker adaptation algorithms converges in the error similar to that using the speaker dependent model. Note that the model topology

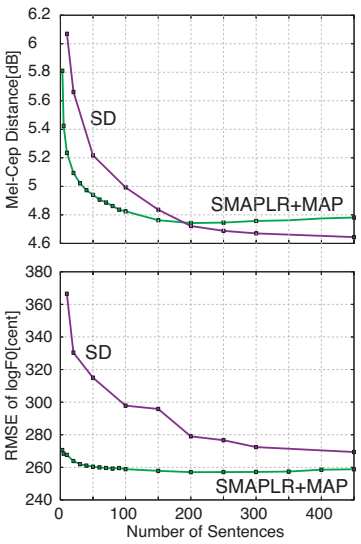

Fig. 2. Objective evaluation of speaker adaptation algorithm and speaker dependent algorithms.

of the average voice model defined by the decision tree is not the same as the speaker dependent model, and the adaptation data includes more or less voice-quality variation. As a result, the speaker adaptation performance does not converge to the speaker-dependent performance.

6.3. Subjective Evaluation of Speaker Adaptation Method and Speaker Dependent Method

We then conducted a Comparison Category Rating (CCR) test to evaluate voice characteristics and prosodic features of synthesized speech using SMAPLR adaptation, the SMAPLR adaptation and the MAP Modification (SMAPLR+MAP), and speaker dependent model (SD). Seven subjects were first presented reference speech and then synthesized speech samples generated from the models in random order. The subjects were then asked to rate their voice characteristics and prosodic features comparing to those of the reference speech. The reference speech was synthesized by a mel-cepstral vocoder. The rating was done using 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 from 50 test sentences, which were contained in neither training nor adaptation data. Figure 3 shows the results of the CCR test. A confidence interval of 95 % is also shown in the figure. These results confirm again that synthesized speech of the speaker adaptation algorithm (SMAPLR+MAP) significantly outperforms that of the speaker dependent model when the adaptation data is relatively limited, and furthermore, when relatively sufficient amount of the adaptation data is available, both synthetic speech have almost the same score.

7. CONCLUSIONS

This paper has described several HSMM-based speaker adaptation algorithms and MAP modification algorithm to develop consistent method for synthesizing speech in a unified way for arbitrary amount of the speech data. From the results of subjective and objective evaluation tests, we have evaluated and shown the advantages of the speaker adaptation algorithms and MAP modification. Our future work is integration of SMAPLR and CMLLR adaptation.

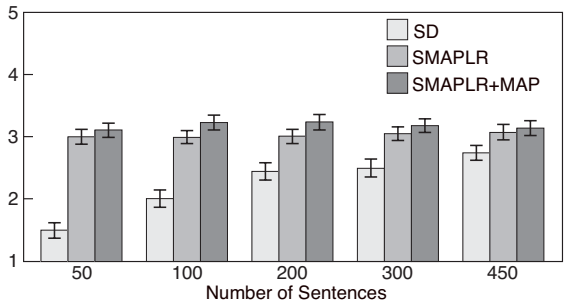

Fig. 3. Subjective evaluation of speaker adaptation algorithm and speaker dependent algorithms.

8. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[6] M. Rahim and B.H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 19–30, Jan. 1996.

[7] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure," in *Proc. EUROSPEECH-95*, Sept. 1995, pp. 1143–1146.

[8] K. Shinoda and C.H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 276–287, Mar. 2001.

[9] 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.

[10] M.J.F. Gales, "Multiple-cluster adaptive training schemes," in *Proc. ICASSP 2001*, May 2001, pp. 361–364.

[11] 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.

[12] O. Shiohan, T.A. Myrvoll, and C-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, no. 3, pp. 5–24, 2002.

[13] V. Digalakis and L.Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 294–300, July 1996.