Acoustic Model Training Based on Linear Transformation and MAP Modification for HSMM-Based Speech Synthesis

Katsumi Ogata, Makoto Tachibana, Junichi Yamagishi, Takao Kobayashi

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

10.21437/Interspeech.2006-389

Abstract

This paper describes the use of combined linear regression and *ex-post* MAP methods for average-voice-based speech synthesis system based on HMM. To generate more natural sounding speech using the average-voice-based speech synthesis system when a large amount of training data is available, we apply *ex-post* MAP estimation after the linear transformation based adaptation. We investigate how the amount of data used in the training of the average voice model and the tying topology affect the naturalness of synthetic speech. From the results of evaluation tests, we show that the adapted average voice model trained using a large amount of data can generate more natural sounding speech than the speaker dependent model.

Index Terms: HMM-based speech synthesis, HSMM, average voice model, speaker adaptation.

1. Introduction

For the purpose of achieving speech synthesis with an arbitrary speaker's voice, we have proposed a statistical speech synthesis approach based on *average voice model* and speaker adaptation [1],[2]. Using a speaker adaptation algorithm based on linear transformation, this approach enables us to synthesize more natural sounding speech than a method based on speaker dependent (SD) model when limited speech data of the target speaker is available. However, if a large amount of data is available, the average-voice-based approach might provide less natural sounding speech than the method based on the SD model. One of reasons for this is the use of the linear transformation in the speaker adaptation. Specifically, there is an assumption that the target speaker's model is expressed by linear regression of the average voice model. However, this assumption is not always appropriate and it would cause model errors.

To overcome this problem, we incorporate a combined approach [3], or *ex-post* maximum a posteriori (MAP) estimation, into the average-voice-based speech synthesis technique [4]. More specifically, after adapting the average voice model to the target speaker using a linear transformation-based algorithm, we further apply MAP estimation as shown in Fig. 1. The MAP estimation theoretically approaches ML estimation, which is used for the training of the SD model, as the amount of data increases. In the average voice model, speech units are modeled using HMMs whose states are tied, and the tying topology of the model parameters depends on the amount of data which is used for training of the average voice model. As a result, the MAP estimation does not approach the ML estimation of the SD model due to the difference of tying topology.

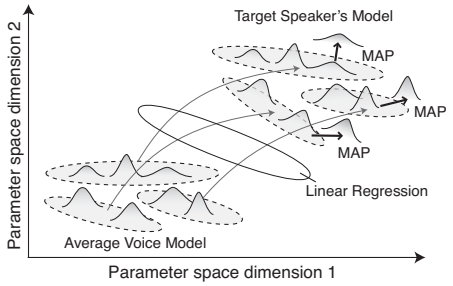

Figure 1: *Modification by the Maximum a posteriori*.

In this paper, we first show experimentally that the *ex-post* MAP estimation approaches the ML estimation asymptotically when the average voice model is adapted using the same tying topology. Then, we investigate how the amount of data used in the training of the average voice model and the tying topology affect the naturalness of synthetic speech. Furthermore, we show superiority of the average-voiced-based technique to the technique based on the SD model from the results of subjective and objective evaluation tests.

2. Relationship between MAP and ML Estimates

In speaker adaptation for speech synthesis, to convert spectrum, fundamental frequency (F0), and phone duration appropriately, we utilize a framework of hidden semi-Markov model (HSMM) [4],[5] which is an HMM with explicit state duration probabilities instead of the transition probabilities. In this paper, we assume that state output and duration distributions are given by Gaussian density functions as follows:

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

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

where $\boldsymbol{o}, \boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are observation vector, mean vector and covariance matrix of output distribution, and d, m_i, and σ_i^2 are state duration, mean and variance of state duration distribution, respectively.

Here we consider the problem that average voice model and adaptation data are given. Let $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ be the adaptation data of length T. We assume that the mean vectors $\overline{\boldsymbol{\mu}}_i$ and

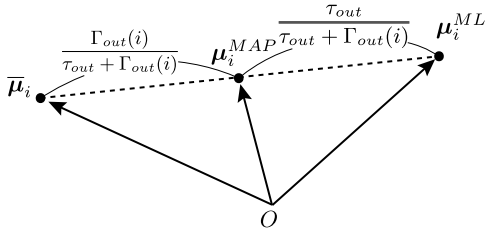

Figure 2: *Relationship between MAP and ML estimates*.

$\overline{m}_i$ are obtained after linear regression based adaptation using the adaptation data. If we estimate the mean vectors based on EM algorithm by using the same adaptation data, we obtain ML estimation as follows:

$$\boldsymbol{\mu}_i^{ML} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{o}_s}{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d} \tag{3}$$

$$m_i^{ML} = \frac{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d}{\sum_{t=1}^{T}\sum_{d=1^t}\gamma_t^d(i)} \tag{4}$$

where $\gamma_t^d(i)$ is the probability generating serial observation sequence $\boldsymbol{o}_{t-d+1}, \ldots, \boldsymbol{o}_t$ at the i-th state.

In contrast, *ex-post* MAP estimation can be applied to the adapted model using the adaptation data. Then, MAP estimation is derived as follows:

$$\boldsymbol{\mu}_i^{MAP} = \frac{\tau_{out}\overline{\boldsymbol{\mu}}_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{o}_s}{\tau_{out}+\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d} \tag{5}$$

$$m_i^{MAP} = \frac{\tau_{dur}\overline{m}_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d}{\tau_{dur}+\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{6}$$

where $\overline{\boldsymbol{\mu}}_i$ and $\overline{m}_i$ are the mean vectors transformed by the linear regression, and τ_{out} and τ_{dur} are positive hyper-parameters of the MAP estimation [6] for the state output and duration distributions, respectively. We can rewrite (5) and (6) as

$$\boldsymbol{\mu}_i^{MAP} = \frac{\tau_{out}}{\tau_{out}+\Gamma_{out}(i)}\overline{\boldsymbol{\mu}}_i + \frac{\Gamma_{out}(i)}{\tau_{out}+\Gamma_{out}(i)}\boldsymbol{\mu}_i^{ML} \tag{7}$$

$$m_i^{MAP} = \frac{\tau_{dur}}{\tau_{dur}+\Gamma_{dur}(i)}\overline{m}_i + \frac{\Gamma_{dur}(i)}{\tau_{dur}+\Gamma_{dur}(i)}m_i^{ML} \tag{8}$$

where

$$\Gamma_{out}(i) = \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d \tag{9}$$

$$\Gamma_{dur}(i) = \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i). \tag{10}$$

As a result, the MAP estimate $\boldsymbol{\mu}_i^{MAP}$ can be viewed as a weighted average of the adapted mean vector $\overline{\boldsymbol{\mu}}_i$ and the ML-estimated mean vectors $\boldsymbol{\mu}_i^{ML}$ as shown in Fig. 2. Similarly, m_i^{MAP} can be viewed as that of $\overline{m}_i$ and m_i^{ML}. When $\Gamma_{out}(i)$ and $\Gamma_{dur}(i)$ are equal to zero, i.e., no training sample is available, the MAP estimates become $\overline{\boldsymbol{\mu}}_i$ and $\overline{m}_i$. On the other hand, when a large number of training samples are used, i.e., $\Gamma_{out}(i) \to \infty$ or $\Gamma_{dur}(i) \to \infty$, the MAP estimates approach the ML estimates $\boldsymbol{\mu}_i^{ML}$ and m_i^{ML} asymptotically.

3. Experiments

3.1. Experimental Conditions

We used the HSMM-based speech synthesis system described in [2],[5]. Speech database used in the following experiments contained six male and five female speakers' speech samples. Each speaker uttered a set of ATR 503 phonetically balanced sentences. Six male and four female speakers' utterances were taken from the ATR Japanese speech database (Set B) and one female speaker's utterances were taken from neutral reading speech used in [7]. In the modeling of synthesis units, 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. The feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of F0 (logF0), and their delta and delta-delta coefficients. We used 5-state left-to-right HSMMs without skip path. As the speaker adaptation algorithm based on linear transformation, we used structural maximum a posteriori linear regression (SMAPLR) [8]. Furthermore, we applied the *ex-post* MAP estimation to the linear transformed model (SMAPLR+MAP). Fifty test sentences were used for evaluation, which were included in neither training nor adaptation data.

3.2. Objective Evaluation in Same Tying Topology

We first conducted an objective evaluation test for the synthesized speech. In the training stage of the average voice model, STC and speaker adaptive training (SAT) [1],[2] were applied. We chose a male speaker MTK as the target speaker. The average voice model was trained using six male and four female speakers' data including MTK. We constructed the same decision tree common to the training speakers and the target speaker, and used the obtained decision tree as the tying topology of the speaker adaptation of SMAPLR+MAP and ML estimation. By doing this, we can eliminate the influence of the tying topology on the adaptation performance of SMAPLR+MAP and ML estimation. The average voice model was trained using 450 sentences for each speaker, 4500 sentences in total. The adaptation data used in SMAPLR+MAP were 450 sentences of the target speaker included in the training data and the same sentences were used in the ML estimation.

We calculated the target speaker's average mel-cepstral distance and root-mean-square (RMS) error of logF0. In the distance calculation, silence and pause intervals were eliminated. Figure 3 shows the target speaker's average mel-cepstral distance between spectra generated from each model and those obtained by analyzing target speaker's real utterance, and the RMS logF0 error between F0 patterns of synthesized and real speech. The horizontal axis represents the number of adaptation sentences for SMAPLR+MAP. It can be seen that both the mel-cepstral distance and the RMS error of logF0 of SMAPLR+MAP converge to the almost same values as those of the ML estimation. From this result, we can see that the MAP estimation approaches asymptotically the ML estimation by using the same tying topology.

3.3. Objective Evaluation in Different Tying Topology

In general, the amount of training data affects the tying topology of the average voice model. Thus we evaluated here the influence of the amount of training data on the average voice model by an objective evaluation test for synthesized speech. We chose five males and four females as the training speakers for the average voice model, and chose MTK and FTK as the target speakers who were

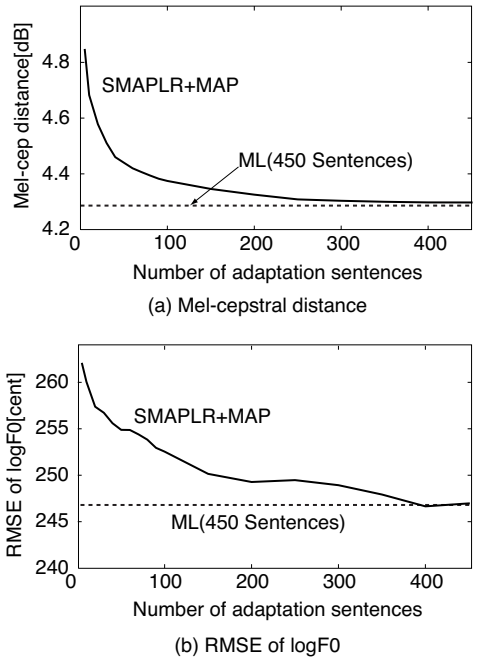

Figure 3: *Objective evaluation of SMAPLR+MAP estimation and ML estimation using same tying topology.*

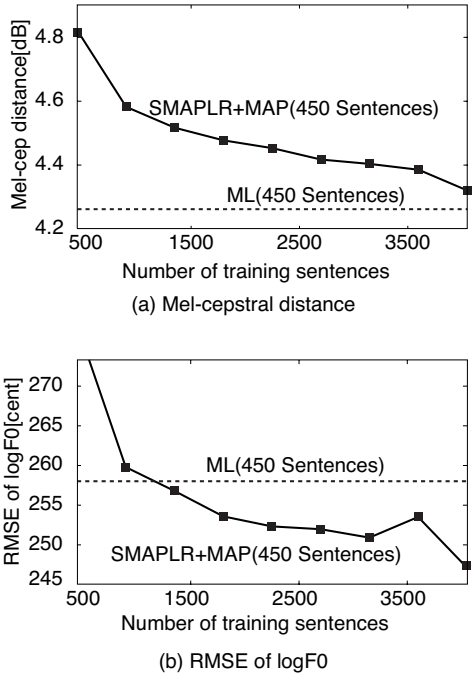

Figure 4: *Objective evaluation of the target speaker MTK.*

not included in the training speakers. The training method of the average voice model was the same as **3.2**. The average voice models were trained using from 50 to 450 sentences for each speaker with increments of 50 sentences, that is, from 450 to 4050 sentences in total. We used different sentence set for each speaker as the training data to avoid dependency on the context. Adaptation data was a set of 450 sentences of the target speaker. We also trained the SD models by the ML estimation using the same 450 sentence set of the target speakers. Note that the average voice models and the SD models have their own tying topology of the model parameters.

Figures 4 and 5 are the results of the objective test. From these figures, it can be seen that both the mel-cepstral distance and RMS error of logF0 of synthesized speech generated from the adapted model become closer to the target speakers' as the the number of training data for average voice model increases. Especially, when the number of training sentences of the average voice model is more than 1350, the adapted model gives better results than the SD model on the RMS logF0 error as shown in Fig. 4. Moreover, we can see that adapted model is closer to the target speaker's logF0 distance than the SD model even when the number of training sentences is 450 in Fig. 5.

3.4. Subjective Evaluation in Different Tying Topology

We finally conducted a comparison category rating (CCR) test to evaluate the naturalness of the synthesized speech using each average voice model and SD model. The average voice models and SD models used in this test were the same as **3.3**. The subjects were first presented with synthesized speech generated from the SD model as reference, then presented with a speech sample chosen randomly from generated speech using the adapted average voice models. The subjects were then asked to rate its naturalness comparing to that of the reference speech. The rating was done using a five-point scale, that is, -2 for much more natural, 0 for almost the same, and +2 for much less natural. For each subject, eight test sentences were randomly chosen from 50 test sentences, which are contained in neither training nor adaptation data. Figure 6 show the results of the CCR test. From this figure, we can see that the scores of synthesized speech using the average voice models are higher than that of using the SD model when the number of training sentences is more than 1350. This result is consistent with the fact in objective test that the adapted model gives better results than the SD model on the RMS logF0 error when the number of training sentences of the average voice model is more than 1350.

3.5. Discussion

From the subjective and objective evaluation tests described in the above, we have seen that we can synthesize more natural sounding speech and closer to the target speaker's feature of logF0 by increasing the training data for the average voice model. One reason for this is the difference of the tying topology of model parameters. To see this, we show the the number of leaf nodes of decision tree in Fig. 7 when changing the number of training sentences for the average voice model. In the figure, we also show the number of leaf nodes of decision tree which is used in the SD model. The SD model was trained using 450 sentences of the target speaker. We can see that the number of leaf nodes increases with increasing the number of training data for the average voice model, and the tying topology of model parameters becomes larger and more complex. Therefore, it is thought that the average voice model can reflect more information in the tying topology than the SD model.

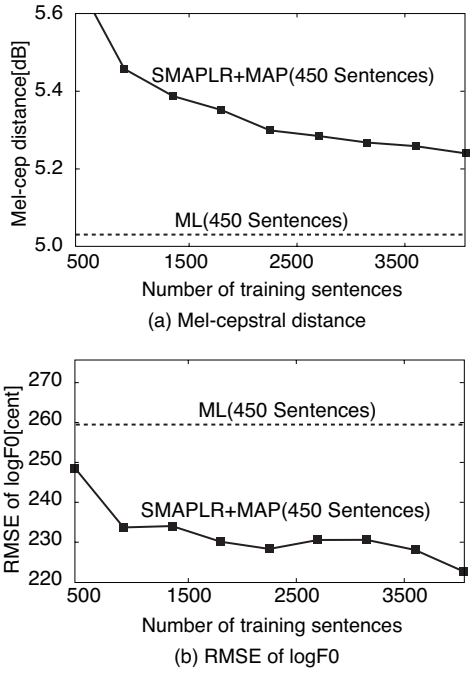

Figure 5: *Objective evaluation of the target speaker FTK.*

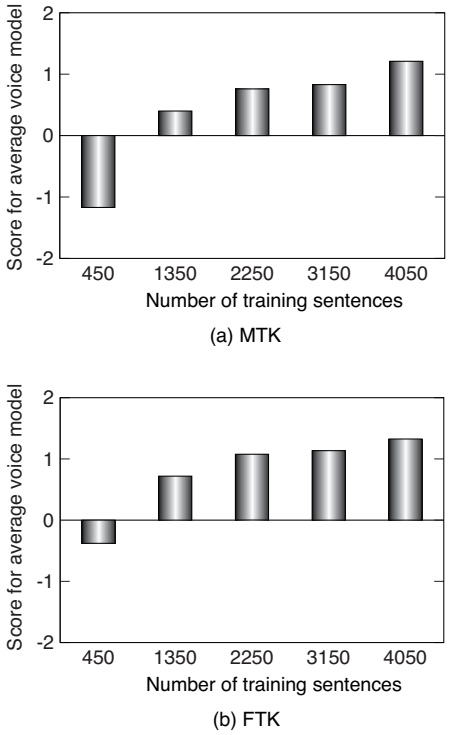

Figure 6: *Subjective evaluation of synthesized voice.*

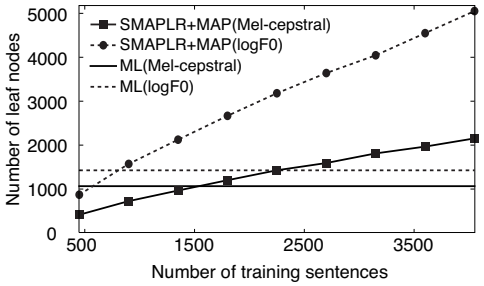

Figure 7: *Number of leaf nodes correspond to each average model.*

4. Conclusions

We have shown the effectiveness of using combined linear regression and MAP modification methods for HMM-based speech synthesis. We have also shown that the MAP estimation approaches asymptotically ML estimation by using the same tying topology. Furthermore, we have examined the influence of the amount of training data for average voice model on the adaptation performance. From the results of subjective and objective evaluation tests, we have shown that the adapted average voice model trained using a large amount of data can generate more natural sounding speech than the SD model.

5. 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, vol. 1, pp. 365–368.

[3] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 4, pp. 294–230, July 1996.

[4] 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, vol. 1, pp. 77–80.

[5] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. INTERSPEECH2004-ICSLP*, Oct. 2004, vol. 2, pp. 1393–1396.

[6] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[7] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expression in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. 3, no. E88-D, pp. 503–509, Mar. 2005.

[8] O. Shiohan, T.A. Myrvoll, and C-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 5–24, Jan. 2002.