Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis

Yuji Nakano, Makoto Tachibana, Junichi Yamagishi, Takao Kobayashi

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

Abstract

This paper proposes a constrained structural maximum a posteriori linear regression (CSMAPLR) algorithm for further improvement of speaker adaptation performance in HMM-based speech synthesis. In the algorithm, the concept of structural maximum a posteriori (SMAP) adaptation is applied to estimation of transformation matrices of the constrained MLLR (CMLLR), where recursive MAP-based estimation of the transformation matrices from the root node to lower nodes of context decision tree is conducted. We incorporate the algorithm into HSMM-based speech synthesis system and show that CSMAPLR adaptation utilizes both of the advantage of CMLLR and SMAPLR adaptation from the result of objective evaluation test. We also show that CSMAPLR adaptation provides more similar synthetic speech to the target speaker than CMLLR and SMAPLR adaptation from the result of subjective evaluation test.

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

1. Introduction

To mimic an arbitrary target speaker's voice using only a small amount of speech data uttered by the target speaker, we have proposed an HMM-based speech synthesis approach using speaker adaptation and average voice model [1]-[3]. In this approach, first, spectrum, fundamental frequency (F0), and duration of several training speakers are modeled simultaneously in a framework of HMM, and an average voice model, which models average voice and prosodic characteristics of the multi speakers, is trained by using adaptive training for the speaker normalization [2],[3]. Then, using a speaker adaptation algorithm such as maximum likelihood linear regression (MLLR) adaptation [4], the average voice model is adapted to a target speaker using a small amount of speech data uttered by the target speaker. After the speaker adaptation, speech is synthesized in the same way as the speaker-dependent HMM-based speech synthesis method [5],[6].

Furthermore, we have investigated several speaker adaptation algorithms [7] for the average-voice-based speech synthesis. As a result, we see that the constrained MLLR (CMLLR) adaptation [8] and the structural maximum a posteriori linear regression (SMAPLR) adaptation [9] are promising approaches. In this paper, for further improvement of the speaker adaptation, we derive a constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation algorithm that integrates CMLLR adaptation and SMAPLR adaptation. In the proposed algorithm, the concept of structural maximum a posteriori (SMAP) adaptation [10] is applied to the estimation of the transformation matrices of CMLLR, that is, recursive MAP-based estimation of the transformation matrices is conducted from the root node to lower nodes of context decision tree. We show results of objective and subjective evaluation tests and effectiveness of the proposed algorithm.

2. CSMAPLR Algorithm for HSMM-based Speech Synthesis

2.1. Speaker Adaptation Based on HSMM

We briefly describe speaker adaptation algorithms reformulated for hidden semi-Markov model (HSMM) in [7]. We assume that each speech synthesis unit is modeled by an N-state HSMM λ. We also assume that the i-th state output $b_i(\boldsymbol{o})$ and duration distributions $p_i(d)$ 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 vector and d is the time staying in the state i.

In MLLR adaptation [4], which is the most popular linear regression adaptation, the mean vectors of state output and duration distributions for the target speaker are obtained by linearly transforming those of the average voice model. However, the transformation of the MLLR adaptation is applied only to the mean vectors of the initial model. On the other hand, in CMLLR adaptation [8], the mean vectors and covariance matrices of state output and duration distributions for the target speaker are transformed at the same time. Since the range of the variation is one of the important factors for F0 and CMLLR can tune not only mean values but also the ranges of the variation to the target speaker, CMLLR would conduct more appropriate adaptation of prosodic information.

Meanwhile, to determine the tying topology for the transformation matrices, we utilize context decision trees [11] in which questions are related to the suprasegmental features, such as mora, accentual phrase, part of speech, breath group, and sentence information. This is because prosodic feature is characterized by many suprasegmental features. In SMAPLR adaptation [9], the concept of SMAP adaptation [10] is applied to the estimation of the transformation matrices of MLLR, that is, recursive MAP-based estimation of the transformation matrices is conducted from the root node to lower nodes of the context decision tree. As a result, we can make better use of the structural information and the suprasegmental information which the context decision tree has.

2.2. Constrained Structural Maximum A Posteriori Linear Regression

The CMLLR adaptation algorithm utilizes the structural information less effectively than the SMAPLR adaptation, whereas the transformed parameters of the SMAPLR adaptation algorithm is applied only to the mean vectors of the average voice model. Here we apply the concept of the SMAP adaptation to the estimation of the transformation matrices of CMLLR.

In the CSMAPLR adaptation, like the 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:

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}'\boldsymbol{\mu}_i - \boldsymbol{\epsilon}', \boldsymbol{\zeta}'\boldsymbol{\Sigma}_i\boldsymbol{\zeta}'^{\top}) \quad (3)$$
$$= |\boldsymbol{\zeta}|\,\mathcal{N}(\boldsymbol{\zeta}\boldsymbol{o} + \boldsymbol{\epsilon}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (4)$$
$$= |\boldsymbol{\zeta}|\,\mathcal{N}(\boldsymbol{W}\boldsymbol{\xi}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (5)$$
$$p_i(d) = \mathcal{N}(d; \chi' m_i - \nu', \chi'\sigma_i^2\chi') \quad (6)$$
$$= |\chi|\,\mathcal{N}(\chi d + \nu; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (7)$$
$$= |\chi|\,\mathcal{N}(\boldsymbol{X}\boldsymbol{\phi}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (8)$$

where $\boldsymbol{\zeta} = \boldsymbol{\zeta}'^{-1}$, $\boldsymbol{\epsilon} = \boldsymbol{\zeta}'^{-1}\boldsymbol{\epsilon}'$, $\chi = \chi'^{-1}$, and $\nu = \chi'^{-1}\nu'$. $\boldsymbol{\xi} = [\boldsymbol{o}^{\top}, 1]^{\top}$ and $\boldsymbol{\phi} = [d, 1]^{\top}$, and $\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}]$ and $\boldsymbol{X} = [\chi, \nu]$ are the transformation matrices.

Moreover, in CMLLR, ML-based estimation is used for obtaining the transformation matrices, whereas, in CSMAPLR, like the SMAPLR adaptation, MAP-based estimation is used as follows:

$$\overline{\boldsymbol{\Lambda}} = (\overline{\boldsymbol{W}}, \overline{\boldsymbol{X}}) = \underset{\boldsymbol{\Lambda}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\boldsymbol{\Lambda}, \boldsymbol{\lambda})\, P(\boldsymbol{\Lambda}) \quad (9)$$

where $\boldsymbol{W} = \{\boldsymbol{W}_j\}_{j=1}^{M}$, $\boldsymbol{X} = \{\boldsymbol{X}_j\}_{j=1}^{M}$, $\boldsymbol{\Lambda} = (\boldsymbol{W}, \boldsymbol{X})$, and M is the total number of distributions. $P(\boldsymbol{\Lambda})$ is the a priori distribution for the transformation matrix $\boldsymbol{W}$ and $\boldsymbol{X}$.

A convenient prior distribution families for $P(\boldsymbol{W})$ and $P(\boldsymbol{X})$ are the matrix variate normal distributions, matrix versions of the multivariate normal distribution, defined as follows [12]:

$$p(\boldsymbol{W}) \propto |\boldsymbol{\Omega}|^{-(a+1)/2} |\boldsymbol{\Psi}|^{-a/2} \exp\Big\{-\frac{1}{2} tr(\boldsymbol{W} - \boldsymbol{H})^T \boldsymbol{\Omega}^{-1} (\boldsymbol{W} - \boldsymbol{H}) \boldsymbol{\Psi}^{-1}\Big\} \quad (10)$$

$$p(\boldsymbol{X}) \propto |\boldsymbol{\omega}|^{-(b+1)/2} |\boldsymbol{\psi}|^{-b/2} \exp\Big\{-\frac{1}{2} tr(\boldsymbol{X} - \boldsymbol{\eta})^T \boldsymbol{\omega}^{-1} (\boldsymbol{X} - \boldsymbol{\eta}) \boldsymbol{\psi}^{-1}\Big\} \quad (11)$$

where $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}$, $\boldsymbol{H}$, $\boldsymbol{\omega}$, $\boldsymbol{\psi}$, and $\boldsymbol{\eta}$ are the hyperparameters for those distribution families, and a and b are the dimensions of the mean vector $\boldsymbol{\mu}$ and mean m, respectively. $\boldsymbol{H}$ and $\boldsymbol{\eta}$ are the transformation matrices of parents node. The scales of the prior distributions $p(\boldsymbol{W})$ and $p(\boldsymbol{X})$ are controlled by the two hyperparameters $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\omega}$, $\boldsymbol{\psi}$, respectively. We fix $\boldsymbol{\Psi}$ and $\boldsymbol{\psi}$ to the identity matrices, $\boldsymbol{\Psi} = \boldsymbol{I} \in \mathbb{R}^{(a+1)\times(a+1)}$ and $\boldsymbol{\psi} = \boldsymbol{I} \in \mathbb{R}^{(b+1)\times(b+1)}$. $\boldsymbol{\Omega}$ and $\boldsymbol{\omega}$ are set to scaled identity matrices, $\boldsymbol{\Omega} = C \cdot \boldsymbol{I}$ and $\boldsymbol{\omega} = \tau \cdot \boldsymbol{I}$ so that the scaling is only controlled by scalar coefficients $C > 0$ and $\tau > 0$.

Re-estimation formulas based on the Baum-Welch algorithm of the transformation matrices can be derived as follows:

$$\overline{\boldsymbol{w}}_l = (\alpha \boldsymbol{p}_l + \boldsymbol{y}_l) \boldsymbol{G}_l^{-1} \quad (12)$$

$$\overline{\boldsymbol{X}} = (\beta \boldsymbol{q} + \boldsymbol{z}) \boldsymbol{K}^{-1} \quad (13)$$

where $\boldsymbol{w}_l$ is the l-th row vector of $\boldsymbol{W}$, $\boldsymbol{p}_l = [0, \boldsymbol{c}_l]$, $\boldsymbol{q} = [0, 1]$, and $\boldsymbol{c}_l$ is the l-th cofactor row vector of $\boldsymbol{W}$. Then $\boldsymbol{y}_l$, $\boldsymbol{G}_l$, $\boldsymbol{z}$, and $\boldsymbol{K}$ are given by

$$\boldsymbol{y}_l = \sum_{r=1}^{R_b} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\Sigma_r(l)}\, \mu_r(l) \sum_{s=t-d+1}^{t} \boldsymbol{\xi}_s^{\top} + C \cdot \boldsymbol{H}(l) \quad (14)$$

$$\boldsymbol{G}_l = \sum_{r=1}^{R_b} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\Sigma_r(l)} \sum_{s=t-d+1}^{t} \boldsymbol{\xi}_s \boldsymbol{\xi}_s^{\top} + C \cdot \boldsymbol{I} \quad (15)$$

$$\boldsymbol{z} = \sum_{r=1}^{R_p} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\sigma_r^2}\, m_r\, \boldsymbol{\phi}_r^{\top} + \tau \cdot \boldsymbol{\eta} \quad (16)$$

$$\boldsymbol{K} = \sum_{r=1}^{R_p} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\sigma_r^2}\, \boldsymbol{\phi}_r\, \boldsymbol{\phi}_r^{\top} + \tau \cdot \boldsymbol{I} \quad (17)$$

where $\Sigma_r(l)$ is the l-th diagonal element of $\boldsymbol{\Sigma}_r$, $\mu_r(l)$ is the l-th element of the mean vector of $\boldsymbol{\mu}_r$, and $\boldsymbol{H}(l)$ is the l-th row vector of $\boldsymbol{H}$. Note that $\boldsymbol{W}$ and $\boldsymbol{X}$ are tied across R_b and R_p distributions, respectively. Then α and β are scalar values which satisfy the following quadratic equations:

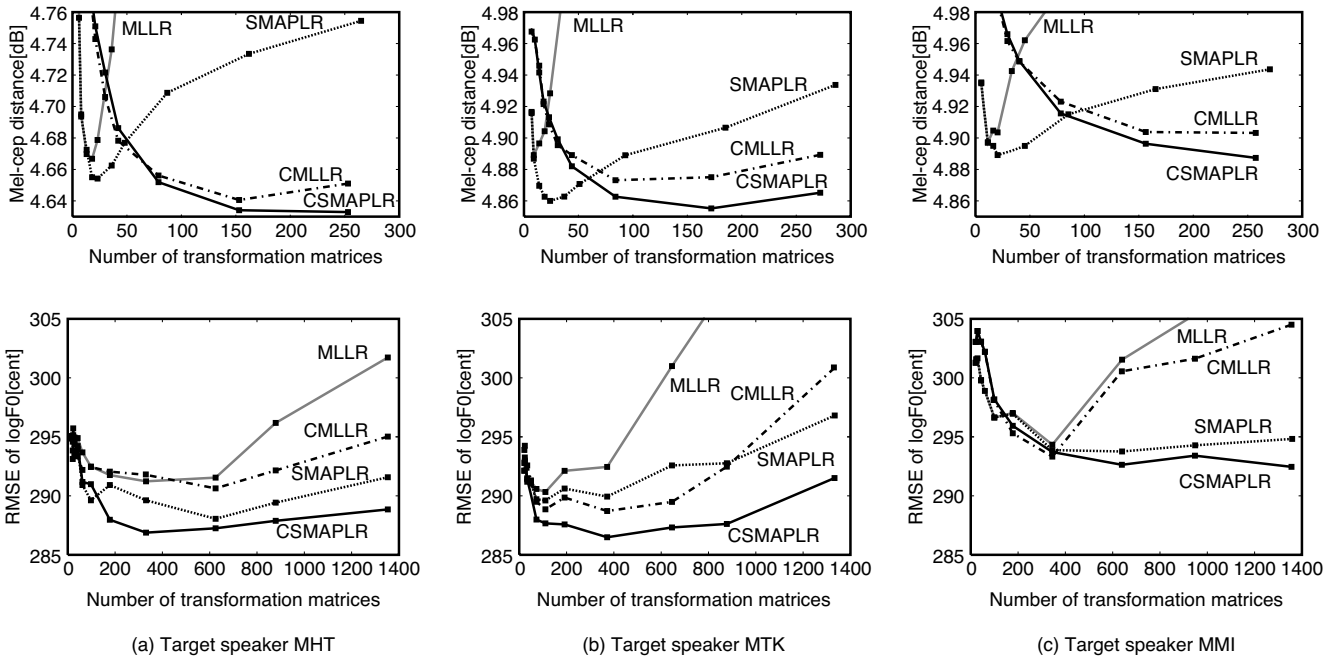
$$\alpha^2 \boldsymbol{p}_l \boldsymbol{G}_l^{-1} \boldsymbol{p}_l^{\top} + \alpha \boldsymbol{p}_l \boldsymbol{G}_l^{-1} \boldsymbol{y}_l^{\top} - \sum_{r=1}^{R_b} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r)\, d = 0 \quad (18)$$

$$\beta^2 \boldsymbol{q} \boldsymbol{K}^{-1} \boldsymbol{q}^{\top} + \beta \boldsymbol{q} \boldsymbol{K}^{-1} \boldsymbol{z}^{\top} - \sum_{r=1}^{R_p} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r) = 0. \quad (19)$$

3. Experiments

3.1. Experimental Conditions

To evaluate the effectiveness of the proposed adaptation algorithm, we conducted objective and subjective evaluation tests for the synthesized speech. Six male and four female speakers' utterances were taken from the ATR Japanese speech database (Set B) and one male speaker's utterances were taken from neutral reading speech used in [13]. Each speaker uttered a set of ATR 503 phonetically balanced sentences. We used 42 phonemes including silence and pause in modeling. 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. We chose four males and four females as the training speakers for the average voice model. A gender-dependent model was trained using four male speakers, 450 sentences for each speakers, 1800 sentences in total. Moreover a gender-independent average voice model was also trained using four male and four female speakers, 450 sentences for each speaker, 3600 sentences in total. Target speakers of the speaker adaptation were the rest of three male speakers MHT, MTK, and MMI who were not used for the training. In the training stage of the average voice models, the shared-decision-tree-based context clustering (STC) algorithm and the speaker adaptive training (SAT) [2],[3] were applied to normalize influence of speaker differences among the training speakers. Note that all the average voice models have the same topology and the number of distributions by using STC. The total number of distributions in each



(a) Target speaker MHT (b) Target speaker MTK (c) Target speaker MMI

Figure 1: *Objective evaluation of speaker adaptation algorithms.*

average voice models was 1861 for spectral part, 2309 for F0 part, and 1121 for phone duration part.

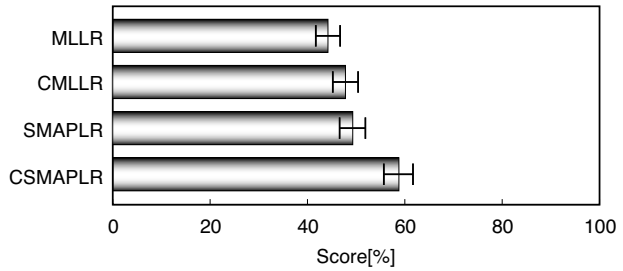
We then adapted the average voice model to the target speaker using fifty adaptation sentences which were not included in the training sentences. The gender-dependent average voice model was used for MHT and MTK, and the gender-independent average voice model was used for MMI. These choices of the average voice models were determined based on a preliminary objective experimental result. The tuning parameters in SMAPLR and CSMAPLR adaptation algorithm, i.e., the thresholds to control hyper-parameters of the MAP estimation, were determined based on a preliminary objective experimental result. The initial prior densities $p(\boldsymbol{W})$ and $p(\boldsymbol{X})$ at the root node were $\boldsymbol{H} = [\boldsymbol{I}, \boldsymbol{0}]$ and $\boldsymbol{\eta} = [1, 0]$, respectively.

3.2. Objective Evaluations of Speaker Adaptation Algorithms

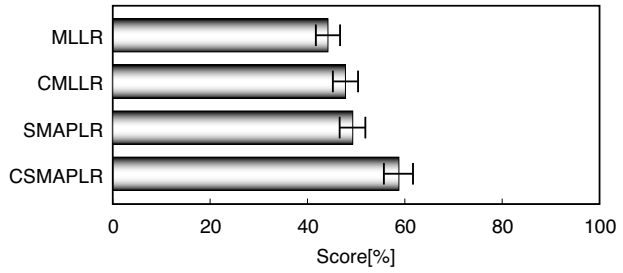
As the objective evaluation for each speaker adaptation algorithm, we calculated the target speaker's mel-cepstral distance between the spectra generated from each model and those obtained by means of analyzing the target speaker's real utterance. In the distance calculation, silence and pause intervals were eliminated. And we calculated the root-mean-square (RMS) error of logF0 between the generated logF0 patterns and those extracted from the target speaker's real utterance. The RMS logF0 error was calculated in the voiced regions only. 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 result of the objective test for three target speakers, MHT, MTK, and MMI. The horizontal axis of the figure represents the number of transformation matrices, which were determined by the thresholds, below which the nodes share the same transformation matrix. As the number of transformation matrices increases, we can use more information of the tree structures. However, when the number of transformation matrices increases too much, the rank-deficient problem occurs in estimation of transformation matrices. And it would decrease estimation accuracy. In MLLR and SMAPLR, when the number of the distribution which shares the transformation matrices is less than that of the dimension of the feature vector, this problem occurs. On the other hand, in CMLLR and CSMAPLR, transformation matrix becomes rank-deficient under the condition that the number of the observation sequences of the adaptation data is less than that of the dimension of the feature vector. Therefore, the rank-deficient problem can occur more easily in MLLR/SMAPLR than CMLLR/CSMAPLR.

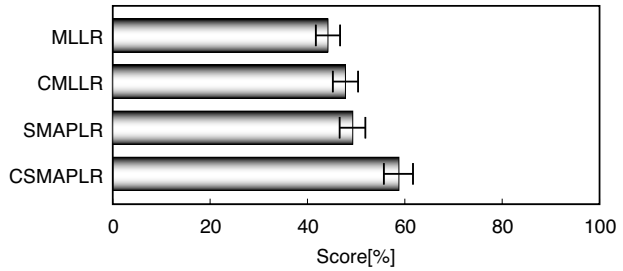
From Fig. 1, it can be seen that the CMLLR adaptation is less sensitive to the change in the number of transformation matrices compared with the MLLR and SMAPLR adaptation. This is due to the above condition for the number of transformation matrices. Moreover, we can see that the CMLLR adaptation gives better results than the MLLR adaptation on the mel-cepstral distance and the RMS logF0 error when the optimal number of the transformation matrices is chosen. This is due to the fact that the CMLLR adaptation can tune not only the mean values but also the ranges of the variation to a new speaker. We can also see that the SMAPLR adaptation gives better results than the MLLR adaptation. This is due to the fact that the SMAPLR adaptation makes better use of the structural information and the suprasegmental information than the MLLR adaptation. Furthermore, we can see that the CSMAPLR adaptation gives better results than the CMLLR and SMAPLR adaptation. As a consequence, we can confirm that the results were improved by using the CSMAPLR adaptation compared with using the CMLLR and SMAPLR adaptation separately.



(a) Target speaker MHT



(b) Target speaker MTK



(c) Target speaker MMI

Figure 2: *Subjective evaluation for voice characteristics of synthesized speech*.

3.3. Subjective Evaluation of Speaker Adaptation Algorithms

We evaluated the voice similarity of the synthesized speech generated from the adapted models by a paired comparison test. Seven subjects were first presented with a reference speech sample, and then a pair of the synthesized speech samples generated from two adapted models chosen from MLLR, CMLLR, SMAPLR, and CSMAPLR in random order. The subjects were then asked which sample was closer to the reference speech. The reference speech was synthesized by a mel-cepstral vocoder. For each subject, eight test sentences were randomly chosen from fifty test sentences, which were contained in neither training nor adaptation data. The thresholds to control the number of transformation matrices were determined based on the result shown in Fig. 1.

Figure 2 shows the preference scores. A confidence interval of 95 % is also shown in the figure. The results confirm that the CSMAPLR adaptation also provides more similar synthetic speech to the target speaker than other adapted models.

4. Conclusions

This paper has proposed CSMAPLR adaptation algorithm which integrated CMLLR and SMAPLR for average-voice-based speech synthesis. From the results of objective and subjective evaluation tests, we have shown that the advantages of the proposed speaker adaptation algorithm. In this paper, variance adaptation was performed only in constrained cases. Thus comparison of constrained method with unconstrained method using the variance adaptation will be our future work. Proposal of a new speaker adaptation algorithm which provides a reduction of dependency on the number of transformation matrices will also be done in the future.

5. References

[1] 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, vol. 1, pp. 345–348.

[2] 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.

[3] J. Yamagishi and T. Kobayashi, "Adaptive training for hidden semi-Markov model," in *Proc. ICASSP 2005*, Mar. 2005, vol. 1, pp. 365–368.

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

[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, vol. 5, pp. 2347–2350.

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

[7] 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.

[8] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[9] 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. 3, pp. 5–24, 2002.

[10] K. Shinoda and C.H. Lee., "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, May 2001.

[11] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *Proc. ICASSP 2004*, May 2004, pp. 5–8.

[12] A.k. Gupta & T. Varga, *Elliptically Contoured Models in Statistics*, Kluwer Academic Publishers, 1993.

[13] 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.