Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm

Junichi Yamagishi, Takao Kobayashi, *Senior Member, IEEE*, Yuji Nakano, Katsumi Ogata, and Juri Isogai

***Abstract*—In this paper, we analyze the effects of several factors and configuration choices encountered during training and model construction when we want to obtain better and more stable adaptation in HMM-based speech synthesis. We then propose a new adaptation algorithm called constrained structural maximum *a posteriori* linear regression (CSMAPLR) whose derivation is based on the knowledge obtained in this analysis and on the results of comparing several conventional adaptation algorithms. Here, we investigate six major aspects of the speaker adaptation: initial models; the amount of the training data for the initial models; the transform functions, estimation criteria, and sensitivity of several linear regression adaptation algorithms; and combination algorithms. Analyzing the effect of the initial model, we compare speaker-dependent models, gender-independent models, and the simultaneous use of the gender-dependent models to single use of the gender-dependent models. Analyzing the effect of the transform functions, we compare the transform function for only mean vectors with that for mean vectors and covariance matrices. Analyzing the effect of the estimation criteria, we compare the ML criterion with a robust estimation criterion called structural MAP. We evaluate the sensitivity of several thresholds for the piecewise linear regression algorithms and take up methods combining MAP adaptation with the linear regression algorithms. We incorporate these adaptation algorithms into our speech synthesis system and present several subjective and objective evaluation results showing the utility and effectiveness of these algorithms in speaker adaptation for HMM-based speech synthesis.**

***Index Terms*—Average voice, hidden Markov model (HMM)-based speech synthesis, speaker adaptation, speech synthesis, voice conversion.**

I. Introduction

RECENT unit-selection and concatenative approaches [1]–[4] produce high-quality synthetic speech, but require large-scale speech corpora if the speech is to sound natural. Using these approaches to develop a humanlike speech synthesizer, which could control many kinds of emotional expressions and speaking styles, we would have to prepare many corpora corresponding to the different styles. These approaches are thus quite unsuitable to the quick addition of several new emotional expressions and speaking styles to a speech synthesizer [5] and are particularly impractical when we need to reproduce intermediate degrees of emotional expressions and speaking styles, such as slightly joyful, somewhat depressed. A statistical parametric speech synthesis system based on hidden Markov models (HMMs) [6]–[12], in contrast, can easily and flexibly generate natural sounding synthetic speech with varying speaking styles and/or emotional expressions. We have indeed already shown that the emotional expressions and speaking styles of synthetic speech can be easily reproduced, controlled, and transformed by using style modeling [13], model adaptation [14], model interpolation and model morphing [15], or multiple-regression HMMs [16].

Another crucial deficiency of speech synthesis systems based on unit selection is the limited number of speakers they can use. Since a unit-selection approach requires to prepare immense corpora corresponding to all the speakers a system uses, we encounter a similar problem to the aforementioned one in a range of emotional expressions and speaking styles when we want to make a speech synthesizer that can simultaneously deal with many speakers' voices. Eliminating this drawback would not only reduce the cost of adding new voices but would also result in many new applications for human–computer interfaces using speech for input and output. For example, it would help personalize speech-to-speech translation so that a user's speech in one language can be used to produce corresponding speech in another language while continuing to sound like the user's voice. To make such speech synthesis with diverse voices and styles feasible, one should minimize the amount of the speech data required for a new speaker, emotional expressions, or speaking style without reducing the quality of the synthetic speech.

For the past ten years, our group has therefore been developing speaker-independent HMM-based speech synthesis in which "average voice models" are created from several speakers' speech data and are adapted with a small amount of speech data from a target speaker (e.g., [17]–[20]). This research started by transforming the spectral parameters of speech [17], [21], [22] by using several speaker adaptation techniques developed for automatic speech recognition such as maximum-likelihood linear regression (MLLR) [23] or MAP-VFS, which is an algorithm combining maximum *a*

Manuscript received October 31, 2007; revised August 01, 2008. Current version published December 11, 2008. This work was supported in part by the JSPS Grant-in-Aid for Scientific Research (B) 15300055, and JSPS Research Fellowships for Young Scientists 164633. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Abeer Alwan.

J. Yamagishi is with the Center for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9LW, U.K. (e-mail: jyamagis@inf.ed.ac.uk).

T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai are with the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan (e-mail: takao.kobayashi@ip.titech.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2006647

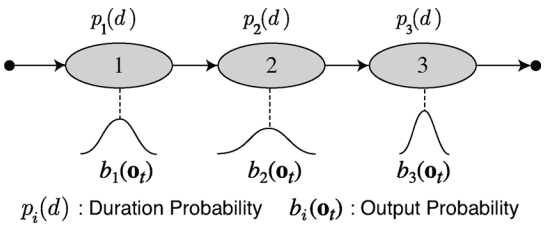


Fig. 1. Hidden semi-Markov model.

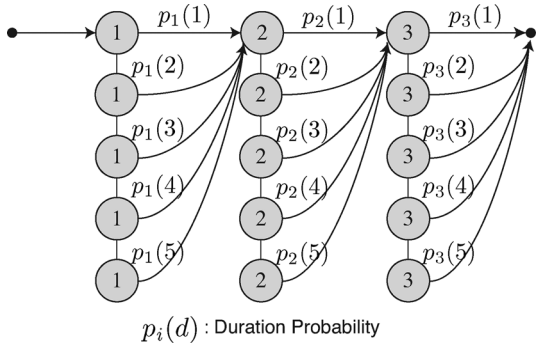


Fig. 2. Duration probability density functions (pdfs) of the hidden semi-Markov model.

posteriori (MAP) adaptation [24] and vector field smoothing (VFS) [25], [26]. Then, to simultaneously model and adapt the excitation parameters of speech as well as the spectral parameters, the multispace probability distribution (MSD) HMM [27] and its MLLR adaptation algorithm [18], [28] have been used. We used the logarithm of the fundamental frequency $(\log F_0)$ and its dynamic and acceleration features as the excitation parameters, and the MSD-HMM enabled us to treat the $\log F_0$ sequence, which is a mixture of one-dimensional real numbers for voiced regions and symbol strings for unvoiced regions, as a probability framework. Furthermore, to also simultaneously model and adapt duration parameters for the spectral and excitation parameters, the MSD hidden semi-Markov model (MSD-HSMM) [29] and its MLLR adaptation algorithm [20] have been used. The HSMM [30]–[32] is an HMM having explicit state duration distributions instead of the transition probabilities to directly model and control phone durations (see Figs. 1 and 2). We also developed several techniques for training the initial model used in these speaker adaptation techniques. The initial model we use is an average voice model constructed from training data which consists of the speech of several speakers. Because this training data includes a lot of speaker-dependent characteristics that affect the adapted models and the quality of synthetic speech generated from them, we incorporated the speaker-adaptive training (SAT) algorithm [33] into our speech synthesis system in order to reduce the negative influence of speaker differences [19]. In the SAT algorithm, the model parameters for the average voice model are obtained using a blind estimation procedure assuming that the speaker difference is expressed by linear transformations of the average voice model. The SAT algorithm for the MSD-HSMM was also derived in [20]. A speaker normalization technique for the tree-based clustering of the model parameters for the average voice model was also developed [34]. Applications to style adaptation (conversion of speaking styles and emotional expressions) and to multilingual/polyglot text-to-speech systems have also been reported [14], [35], [36]. Using this speech synthesis method, which we call "average-voice-based speech synthesis (AVSS)," we can obtain natural-sounding synthetic speech for a target speaker from as little as 100 utterances (about 6 min worth of speech data). Interestingly, we have shown that the speech produced by this approach where the average voice model is trained from enough speech data of several source speakers is perceived as being more natural sounding than that of the speech produced by a speaker-dependent (SD) system using the same 100 utterances or even 450 utterances of the target speaker [20].

For this approach, we have investigated individual adaptation effects of the spectral, excitation, and duration parameters and have investigated the amount of adaptation data required for the adaptation of each parameter in the framework of the MSD-HSMMs [20]. In the previous analysis, we simply applied the MLLR adaptation algorithm having transform functions for only mean vectors to a gender-dependent average voice model. However, many factors on which the performance of the speaker adaptation depends are still unclear in this approach. For example, before we conduct the speaker adaptation, we have to answer question about the following factors.

Initial models: What kinds of average voice model (or speaker-dependent model) is the most appropriate initial model, from which adaptation will start?

The amount of the training data for the initial models: Does the amount of the training data for the initial model affect the adaptation performance? If so, how?

Transform functions: What kinds of transform functions are appropriate? Does the adaptation of covariance matrices affect adaptation performance?

Estimation criteria: Does the robust MAP criterion affect the speaker adaptation for speech synthesis as well as the adaptation for speech recognition?

Sensitivity: In the linear regression algorithms such as MLLR, we need thresholds for controlling the number of the transforms. How do wrong thresholds degrade the adaptation performance? Are there any differences between the thresholds sensitivities of several linear algorithms?

Combination algorithms: Some adaptation algorithms can be combined. Can adaptation performance be improved by combining them?

In this paper, we therefore analyze the ways in which these factors affect HMM-based speech synthesis so that we can find out how to make the speaker adaptation better and more stable. We first analyze the effect of the initial model by comparing the results obtained using 1) speaker-dependent models, 2) gender-independent models, 3) single use of the gender-dependent models, and 4) simultaneous use of the gender-dependent models. We then assess the effect of the transform functions in linear regression algorithms by comparing the transform function for only mean vectors with that for mean vectors and covariance matrices. That is, we compare the transform

TABLE I
DEFINITION OF ACRONYMS FOR FOUR LINEAR REGRESSION ALGORITHMS

	Transform functions	
Estimation criteria	Mean	Mean & covariance
Maximum likelihood	MLLR [23]	CMLLR [37], [38]
Structural MAP	SMAPLR [39]	CSMAPLR

function for MLLR with that for constrained MLLR (CMLLR) [37], [38]. We then analyze the effect of the estimation criteria by comparing the ML criterion with a robust estimation criterion called structural MAP (SMAP) [39] in the linear regression adaptation algorithms (i.e., SMAPLR adaptation [40]). Classification and definition of the acronyms for these linear regression algorithms are shown in Table I. At the same time, we propose and evaluate a new adaptation algorithm, called constrained SMAPLR (CSMAPLR), whose formulation is based on the results of these analyses and a comparison of the results of several conventional adaptation algorithms. We evaluate the threshold sensitivities of four linear regression algorithms described above. As one of the combination algorithms, we choose a method [41] combining MAP adaptation with the linear regression algorithms. We incorporate these adaptation algorithms into our speech synthesis system and show their effectiveness from several subjective and objective evaluation results including our past reports [42]–[46].

This paper is organized as follows. Section II gives an overview of the AVSS system used in our experiments, and in Section III the CMLLR algorithm as well as the new CSMAPLR adaptation algorithm are described in the framework of the HSMM. Section IV-A describes the conditions of the subjective and objective experiments, Section IV-B describes the evaluations of the initial models, Section IV-C describes the evaluation of the transform functions and estimation criteria in these linear regression algorithms, and Section IV-D describes the evaluation of the threshold sensitivities of these algorithms. Section IV-E describes the evaluation of combination algorithms, and Section IV-F, describes the evaluation of the amount of the training data for the average voice model. Section V concludes the paper by briefly summarizing our findings.

II. OVERVIEW OF THE AVSS SYSTEM

As shown in Fig. 3, the average-voice-based speech synthesis system comprises speech analysis, training of the average voice model, speaker adaptation, and speech synthesis.

A. Speech Analysis

From a multi-speaker speech corpus we extract two kinds of parameters required for the Mel-cepstral vocoder with simple pulse or noise excitation: the Mel-cepstral coefficients and $\log F_0$. The Mel-cepstral coefficients are obtained by Mel-cepstral analysis [47]–[49] and the F_0 values are estimated using an instantaneous-frequency-amplitude-spectrum (IFAS)-based method [50]. We use not only these static features but also dynamic and acceleration features. These dynamic and acceleration feature vectors are the first and second delta parameter

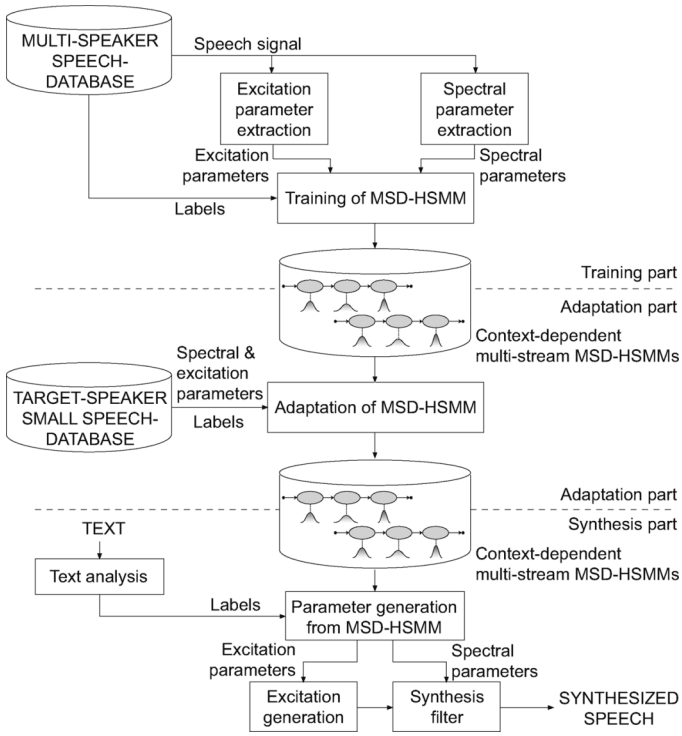


Fig. 3. Overview of the average-voice-based speech synthesis system.

vectors corresponding to the first and second time derivatives of the static feature vector.

B. Acoustic Models and Labels

To model the extracted acoustic features together with their duration in a unified modeling framework, we use context-dependent multi-stream MSD-HSMMs as acoustic units for speech synthesis. The multi-stream model structure is used to model the Mel-cepstral coefficients and $\log F_0$ simultaneously. Japanese phonetic and linguistic contexts used in the following experiments contain phonetic features, mora-level features, morpheme features, accentual features, breath-group-level features, and utterance-level features. Details of the Japanese contexts are as follows [19]:

- preceding, current, and succeeding phonemes;
- the part of speech of the preceding, current, and succeeding morphemes;
- the number of morae and the type of accent in the preceding, current, and succeeding accentual phrases;
- the position of the current mora in the current accentual phrase;
- the differences between the position of the current mora and the type of accent;
- the number of morae in the preceding, current, and succeeding breath groups;
- the position of the current accentual phrase in the current breath group;
- the number of morae in the sentence;
- the position of the breath group in the sentence.

Note that phoneme boundary labels are used only to obtain the initial parameters of the average voice model and that we do not require these labels at the adaptation or synthesis stage.

C. Training of Average Voice Model

Using the above MSD-HSMMs, we train the average voice model as the initial model, from which adaptation will start, from training data that consists of the speech of several speakers. To construct an appropriate average voice model, we use a model-space SAT algorithm [20] to estimate the model parameters and use a shared-decision-tree-based clustering algorithm [34] to tie those parameters.

D. Speaker Adaptation

Using speaker adaptation techniques for the multi-stream MSD-HSMM, we adapt the average voice model to that of the target speaker by using a small amount of speech data. In our conventional system, we used the MLLR adaptation algorithm. In the experiments reported here, we used several linear regression algorithms as shown in Table I.

Like the MLLR adaptation algorithm, the above adaptation algorithms can utilize piecewise linear regression functions. For automatic determination of multiple regression classes for the piecewise linear regression, we used decision trees constructed for the tying of the parameters of the average voice model because the decision trees have phonetic and linguistic contextual questions related to the suprasegmental features by which prosodic features, especially F_0, are characterized [14].

E. Speech Synthesis

In the synthesis step, input text is first transformed into a sequence of context-dependent phoneme labels. A sentence MSD-HSMM corresponding to the label sequence is then constructed by concatenating the context-dependent MSD-HSMMs. Then the Mel-cepstrum and $\log F_0$ corresponding to input text are statistically generated from the sentence MSD-HSMM itself. Here, the duration pdfs automatically determine the duration of each state of the sentence MSD-HSMM. In this system, we use a parameter generation algorithm that uses a maximum-likelihood criterion [6][1] Finally, speech is resynthesized from the generated Mel-cepstral and F_0 parameter sequences by using a Mel-logarithmic spectrum approximation (MLSA) filter.

III. Hidden Semi-Markov Model and Its Speaker Adaptation Techniques

As described in the preceding section, we use the MSD-HSMM framework for the simultaneous transformation of the Mel-cepstrum, $\log F_0$, and duration parameters of speech. For notational simplicity, here we explain the CMLLR adaptation algorithm and the new adaptation algorithm (CSMAPLR) adaptation in the framework of the original HSMM. Extending those algorithms to the MSD-HSMM is straightforward [18].

An N-state left-to-right HSMM λ with no skip paths is specified by a state output probability distribution $\{b_i(\cdot)\}_{i=1}^N$ and a state duration probability distribution $\{p_i(\cdot)\}_{i=1}^N$. We assume that the ith state output and duration distributions are Gaussian distributions respectively characterized by a mean vector $\boldsymbol{\mu}_i \in \mathcal{R}^L$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i \in \mathcal{R}^{L\times L}$ and by a scalar mean m_i and variance σ_i^2. That is

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

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

where $\boldsymbol{o} \in \mathcal{R}^L$ is an L-dimensional observation vector and d is the duration of state i. For a given HSMM λ, the observation probability of training data $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ of length T can be written as

$$P(\boldsymbol{O}|\lambda) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{d=1}^{t} \alpha_{t-d}(j) p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s) \beta_t(i) \tag{3}$$

where $t \in [1, T]$. Then $\alpha_t(i)$ and $\beta_t(i)$ are the forward and backward probabilities defined by

$$\alpha_t(i) = \sum_{d=1}^{t} \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_{t-d}(j) p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s) \tag{4}$$

$$\beta_t(i) = \sum_{d=1}^{T-t} \sum_{\substack{j=1 \\ j\neq i}}^{N} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta_{t+d}(j) \tag{5}$$

where $\alpha_0(i) = 1$ and $\beta_T(i) = 1$. The state occupancy probability $\gamma_t^d(i)$ of being in the state i at the period of time from $t-d+1$ to t is defined as

$$\gamma_t^d(i) = \frac{1}{P(\boldsymbol{O}|\lambda)} \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_{t-d}(j) p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s) \beta_t(i). \tag{6}$$

For further explanation of the training, estimation, and implementation and issues of the HSMMs, see [14], [29], and [20].

A. Constrained Maximum-Likelihood Linear Regression (CMLLR)

Target parameters for the HSMM-based MLLR adaptation that were used in the conventional systems [14], [20] were restricted to the mean vectors of the output and duration pdfs of (1) and (2). In the adaptation for speech synthesis, however, we should adapt both the mean vectors and covariance matrices of the output and duration pdfs to a new speaker because the covariance is also an important factor affecting the characteristics of synthetic speech. In the HMM-based CMLLR adaptation [37], [38], mean vectors and covariance matrices of the state output pdfs are transformed simultaneously using the same matrix (Fig. 4).[2] Similarly, the HSMM-based CMLLR adaptation transforms the mean vectors and covariance matrices of the state output and duration pdfs simultaneously as follows:

$$b_i(\boldsymbol{o}) = \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\zeta}'\boldsymbol{\mu}_i - \boldsymbol{\epsilon}', \boldsymbol{\zeta}'\boldsymbol{\Sigma}_i\boldsymbol{\zeta}'^{\top}\right) \tag{7}$$

$$p_i(d) = \mathcal{N}\left(d; \chi' m_i - \nu', \chi'\sigma_i^2\chi'\right). \tag{8}$$

[1]Experimental results using a recent parameter generation algorithm that considered the global variance [51] were reported in [52], [53].

[2]An adaptation algorithm which transforms mean vectors and covariance matrices of the state output pdfs using different matrices was also proposed [54]. Then it would be possible to apply the SMAP criterion to the adaptation algorithm in a similar way to derive the CSMAPLR adaptation. In this paper, however, we do not consider the *unconstrained* MLLR algorithm and *unconstrained* SMAPLR algorithm since experiments become increasingly complex.

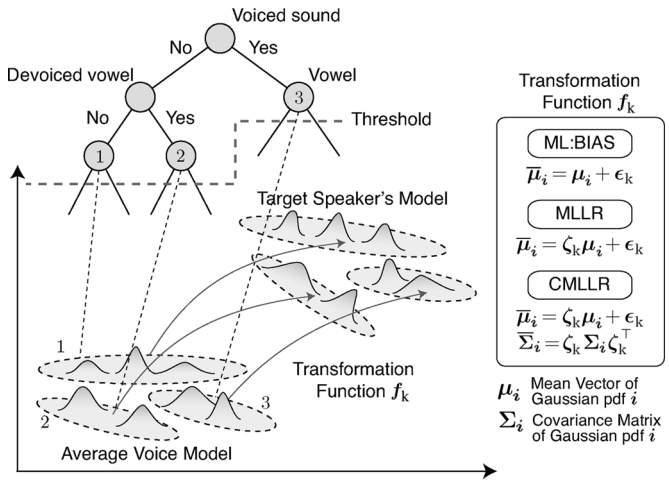


Fig. 4. Constrained maximum-likelihood linear regression (CMLLR) and its related algorithms.

A matrix $\boldsymbol{\zeta}' \in \mathcal{R}^{L\times L}$ is used to transform both the mean vectors and covariance matrices of the state output pdfs and then a scalar χ' is used to transform those of the state duration pdfs. $\boldsymbol{\epsilon}' \in \mathcal{R}^L$ and ν' are bias terms of the transforms. These model transforms are equivalent to the following affine transforms of the feature vectors $\boldsymbol{o}$ and duration d of state i

$$b_i(\boldsymbol{o}) = \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\zeta}'\boldsymbol{\mu}_i - \boldsymbol{\epsilon}', \boldsymbol{\zeta}'\boldsymbol{\Sigma}_i\boldsymbol{\zeta}'^\top\right) \tag{9}$$
$$= |\boldsymbol{\zeta}|\mathcal{N}(\boldsymbol{\zeta}\boldsymbol{o} + \boldsymbol{\epsilon}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{10}$$
$$= |\boldsymbol{\zeta}|\mathcal{N}(\boldsymbol{W}\boldsymbol{\xi}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{11}$$
$$p_i(d) = \mathcal{N}\left(d; \chi' m_i - \nu', \chi'\sigma_i^2\chi'\right) \tag{12}$$
$$= |\chi|\mathcal{N}\left(\chi d + \nu; m_i, \sigma_i^2\right) \tag{13}$$
$$= |\chi|\mathcal{N}\left(\boldsymbol{X}\boldsymbol{\phi}; m_i, \sigma_i^2\right) \tag{14}$$

where $\boldsymbol{\zeta} = \boldsymbol{\zeta}'^{-1}$, $\boldsymbol{\epsilon} = \boldsymbol{\zeta}'^{-1}\boldsymbol{\epsilon}'$, $\chi = \chi'^{-1}$, $\nu = \chi'^{-1}\nu'$, $\boldsymbol{\xi} = [\boldsymbol{o}^\top, 1]^\top$, and $\boldsymbol{\phi} = [d, 1]^\top$. $\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}] \in \mathcal{R}^{L\times(L+1)}$ and $\boldsymbol{X} = [\chi, \nu] \in \mathcal{R}^{1\times 2}$ are, respectively, the linear transform matrices for the state output and duration pdfs.

We estimate a set of transforms $\Lambda = (\boldsymbol{W}, \boldsymbol{X})$ maximizing the likelihood of the adaptation data $\boldsymbol{O}$ of length T

$$\widetilde{\Lambda} = (\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{X}}) = \arg\max_{\Lambda} P(\boldsymbol{O}|\lambda, \Lambda) \tag{15}$$

where λ is the parameter set of the HSMM. Re-estimation formulas based on the EM algorithm [55] of lth row vector $\boldsymbol{w}_l$ of $\boldsymbol{W}$ and $\boldsymbol{X}$ can then be derived as follows:

$$\widetilde{\boldsymbol{w}}_l = (\alpha\boldsymbol{p}_l + \boldsymbol{y}_l)\boldsymbol{G}_l^{-1} \tag{16}$$
$$\widetilde{\boldsymbol{X}} = (\beta\boldsymbol{q} + \boldsymbol{z})\boldsymbol{K}^{-1} \tag{17}$$

where $\boldsymbol{p}_l = [0\ \boldsymbol{c}_l^\top]^\top$ and $\boldsymbol{q} = [0\ 1]^\top$. Note that $\boldsymbol{c}_l$ is lth cofactor row vector of $\boldsymbol{W}$. The terms $\boldsymbol{y}_l \in \mathcal{R}^{L+1}$, $\boldsymbol{G}_l \in \mathcal{R}^{(L+1)\times(L+1)}$, $\boldsymbol{z} \in \mathcal{R}^2$, and $\boldsymbol{K} \in \mathcal{R}^{2\times 2}$ in these equations are given by

$$\boldsymbol{y}_l = \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\Sigma_r(l)}\mu_r(l)\sum_{s=t-d+1}^{t}\boldsymbol{\xi}_s^\top \tag{18}$$

$$\boldsymbol{G}_l = \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\Sigma_r(l)}\sum_{s=t-d+1}^{t}\boldsymbol{\xi}_s\boldsymbol{\xi}_s^\top \tag{19}$$

$$\boldsymbol{z} = \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\sigma_r^2}m_r\boldsymbol{\phi}_d^\top \tag{20}$$

$$\boldsymbol{K} = \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\frac{1}{\sigma_r^2}\boldsymbol{\phi}_d\boldsymbol{\phi}_d^\top \tag{21}$$

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

$$\alpha^2\boldsymbol{p}_l\boldsymbol{G}_l^{-1}\boldsymbol{p}_l^\top + \alpha\boldsymbol{p}_l\boldsymbol{G}_l^{-1}\boldsymbol{y}_l^\top - \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)d = 0 \tag{22}$$

$$\beta^2\boldsymbol{q}\boldsymbol{K}^{-1}\boldsymbol{q}^\top + \beta\boldsymbol{q}\boldsymbol{K}^{-1}\boldsymbol{z}^\top - \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r) = 0. \tag{23}$$

Since the cofactor $\boldsymbol{c}_l$ affects all row vectors of $\widetilde{\boldsymbol{W}}$, we update $\widetilde{\boldsymbol{W}}$ using an iterative method proposed in [8]. On the other hand, the estimation of $\widetilde{\boldsymbol{X}}$ in (17) has a closed-form solution. Although we explain this algorithm using a global transform, the explanation can be straightforwardly extended to estimating multiple transforms and conducting piecewise linear regression. To group the distributions in the model and to tie the transforms in each group, we use decision trees for context clustering as shown in Fig. 4.

This algorithm would have an effect on adaptation of prosodic information because the ranges of F_0 and duration are important factors for synthetic speech. For example, if the target speaker has a speaking style characterized by modulation and rhythm—that is, with F_0 and duration ranges wider than those of the average voice model—we cannot mimic that style without adapting the variance of the average voice model. It could yield a similar benefit in adaptation of the Mel-cepstrum.

Another advantage of the CMLLR adaptation algorithm is that we can efficiently transform the diagonal covariance matrices of the Gaussian distributions of the average voice model into full matrices in the parameter generation algorithm. In the systems using the diagonal covariance matrices, each acoustic feature dimension is optimized independently and thus it sometimes generates an artificial sound [56]. This limitation is then addressed by the use of full-covariance modeling techniques, which is able to reflect within-frame correlations. In [56], it is reported that full covariance modeling using semi-tied covariance [57] (also known as maximum-likelihood linear transformation [58]) has an effect on the parameter-generation algorithm considering global variance (GV) [51]. As we can see from (7), the full covariance can be modeled by using the CMLLR transform instead of the semi-tied covariance.

In addition to these MLLR and CMLLR adaptation algorithms, single bias removal [59], SMAP adaptation [39], SMAPLR adaptation [40], multiple linear regression called

ESAT [60] and so on can be also defined in the framework of the HSMM [43].

B. Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR)

The CMLLR adaptation algorithm uses the maximum-likelihood criterion for the estimation of the transforms. The criterion would work well in the training stage of the average voice model using the SAT algorithm because a large amount of training data for the average voice model is available. In the adaptation stage, however, the amount of adaptation data is limited and we therefore need to use a more robust criterion, such as the maximum *a posteriori* criterion. In the MAP estimation, we estimate the transforms as follows:

$$\widehat{\Lambda} = (\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{X}}) = \arg\max_{\Lambda} P(\boldsymbol{O}|\lambda, \Lambda) P(\Lambda) \tag{24}$$

where $P(\Lambda)$ is a prior distribution for the transforms $\boldsymbol{W}$ and $\boldsymbol{X}$. For the prior distribution, the following combined matrix variate normal distributions (matrix versions of the multivariate normal distribution [61]) are convenient:

$$\begin{aligned} P(\Lambda) \propto & |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} |\tau_p|^{-1} |\boldsymbol{\psi}|^{-\frac{1}{2}} \\ & \times \exp\left\{ -\frac{1}{2} tr(\boldsymbol{W} - \boldsymbol{H})^{\top} \boldsymbol{\Omega}^{-1} (\boldsymbol{W} - \boldsymbol{H}) \boldsymbol{\Psi}^{-1} \right\} \\ & \times \exp\left\{ -\frac{1}{2} tr(\boldsymbol{X} - \boldsymbol{\eta})^{\top} \tau_p^{-1} (\boldsymbol{X} - \boldsymbol{\eta}) \boldsymbol{\psi}^{-1} \right\} \end{aligned} \tag{25}$$

where $\propto$ means proportion, $\boldsymbol{\Omega} \in \mathcal{R}^{L \times L}$, $\boldsymbol{\Psi} \in \mathcal{R}^{(L+1)\times(L+1)}$, $\boldsymbol{H} \in \mathcal{R}^{L \times (L+1)}$, $\tau_p > 0$, $\boldsymbol{\psi} \in \mathcal{R}^{2\times 2}$, and $\boldsymbol{\eta} \in \mathcal{R}^{1\times 2}$ are the hyperparameters for the prior distribution.

In the SMAP criterion [39], tree structures of the distributions effectively cope with the control of the hyperparameters. Specifically, we first use all the adaptation data to estimate a global transform at the root node of the tree structure, and then propagate it to its child nodes as their hyperparameters $\boldsymbol{H}$ and $\boldsymbol{\eta}$. In the child nodes, their transforms are estimated again using their adaptation data and using the MAP criterion with the propagated hyperparameters. Then the recursive MAP-based estimation of the transforms from the root node to lower nodes is conducted (Fig. 5). Shiohan *et al.* developed SMAPLR adaptation [40] by applying the SMAP criterion to MLLR.

In this paper, we apply the SMAP criterion to the CMLLR adaptation and use the recursive MAP criterion to estimate the transforms for simultaneously transforming the mean vectors and covariance matrices of state output and duration distributions. This algorithm is called constrained structural maximum *a posteriori* linear regression (CSMAPLR). In CSMAPLR adaptation, we fix $\boldsymbol{\Psi}$ and $\boldsymbol{\psi}$ to the identity matrices and set $\boldsymbol{\Omega}$ to a scaled identity matrix $\boldsymbol{\Omega} = \tau_b \boldsymbol{I}_L$ so that the scaling is controlled by a positive scalar coefficient τ_b in the same manner as in SMAPLR adaptation [40]. Here $\boldsymbol{I}_L$ is the $L \times L$ identity matrix. We use the same notation method for different-dimensional identity matrices. Re-estimation formulas based on the EM algorithm [55] of the transforms can be derived as follows:

$$\widehat{\boldsymbol{w}}_l = (\alpha \boldsymbol{p}_l + \boldsymbol{y}'_l) \boldsymbol{G}'^{-1}_l \tag{26}$$

$$\widehat{\boldsymbol{X}} = (\beta \boldsymbol{q} + \boldsymbol{z}') \boldsymbol{K}'^{-1} \tag{27}$$

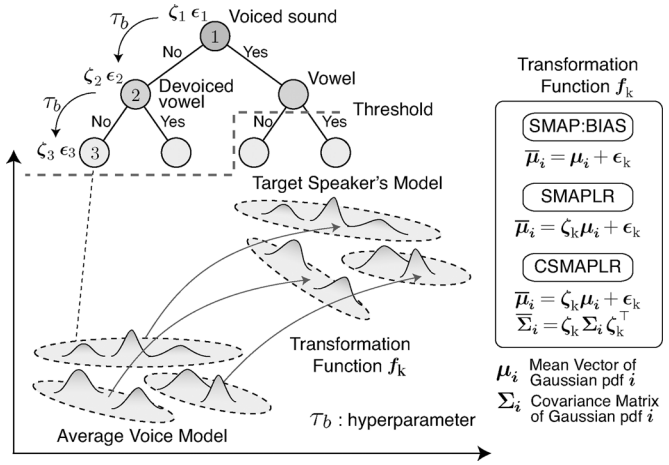


Fig. 5. Constrained structural maximum *a posteriori* linear regression (CSMAPLR) and its related algorithms.

where $\boldsymbol{p}_l$ and $\boldsymbol{q}$ are the same vectors as those of the CMLLR adaptation. Then $\boldsymbol{y}'_l$, $\boldsymbol{G}'_l$, $\boldsymbol{z}'$, and $\boldsymbol{K}'$ are given by

$$\boldsymbol{y}'_l = \boldsymbol{y}_l + \tau_b \boldsymbol{h}_l \tag{28}$$

$$\boldsymbol{G}'_l = \boldsymbol{G}_l + \tau_b \boldsymbol{I}_{L+1} \tag{29}$$

$$\boldsymbol{z}' = \boldsymbol{z} + \tau_p \boldsymbol{\eta} \tag{30}$$

$$\boldsymbol{K}' = \boldsymbol{K} + \tau_p \boldsymbol{I}_2 \tag{31}$$

where $\boldsymbol{h}_l$ is the lth row vector of $\boldsymbol{H}$. The quadratic equations for α and β are the same as (22) and (23).

CSMAPLR adaptation algorithm can utilize the tree structure more effectively than CMLLR adaptation can because the tree structure represents connection and similarity between the distributions, and the propagated prior information automatically reflects the connection and similarity. And since the tree structures we used in these experiments represent linguistic information as shown in Figs. 5, the propagated prior information would reflect the connection and similarity of the distributions of the linguistic information.

Computational costs for the CSMAPLR adaptation are as follows. 1) E step in the EM algorithm: The costs are the same as those for the CMLLR adaptation. 2) M step in the EM algorithm: Compared to the CMLLR adaptation, only simple additional operations (28)–(31) are required to estimate a single transform. When a binary tree structure is used for estimating multiple tranforms, the CSMAPLR adaptation has about twice the computational cost as CMLLR, since the transforms are estimated for each node of the binary tree.

C. Combined Algorithm With Linear Regression and MAP Adaptation

Furthermore, we explain a combined algorithm of linear regression and MAP adaptation [41], [62]. In the previous speaker adaptation using linear regression, there is a rough assumption that the target speaker model would be expressed by the piecewise linear regression of the average voice model. By additionally applying the MAP adaptation to the model transformed by the linear regression, it would be possible to appropriately modify the estimation for the distribution having

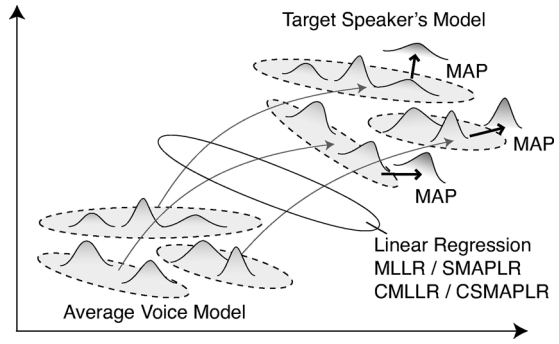


Fig. 6. Combined algorithm of the linear regression and MAP adaptation.

relatively sufficient amount of speech samples (Fig. 6). Here we utilize the same prior distributions as those in [41]. The MAP adaptation of mean vectors of the Gaussian pdfs transformed by the CSMAPLR algorithm can be simply estimated as follows:

$$\widehat{\boldsymbol{\mu}}_i = \frac{\upsilon_b \boldsymbol{\mu}_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\widehat{\boldsymbol{o}}_s}{\upsilon_b + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)d} \tag{32}$$

$$\widehat{m}_i = \frac{\upsilon_p m_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\widehat{d}}{\upsilon_p + \sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)} \tag{33}$$

where $\boldsymbol{\mu}_i$ and m_i are the mean vectors of the state output and duration distributions of the average voice model, and $\widehat{\boldsymbol{o}}_s = \widehat{\boldsymbol{\zeta}}\boldsymbol{o}_s + \widehat{\boldsymbol{\epsilon}}$ and $\widehat{d} = \widehat{\chi}d + \widehat{\nu}$ are linearly transformed observation vector and duration using the HSMM-based CSMAPLR adaptation. Then υ_b and υ_p are positive hyperparameters of the prior distributions for the state output and duration distributions, respectively. Similarly we can combine any linear regression algorithms with this MAP adaptation. As the amount of the adaptation data increases and the number of distributions having relatively sufficient amount of speech samples increases, this algorithm gradually improves the quality of synthetic speech.

IV. Experiments

A. Experimental Conditions

We carried out several subjective and objective evaluation tests to analyze and compare the speaker adaptation algorithms. We used the ATR Japanese speech database (Set B),[3] which contains a set of 503 phonetically balanced sentences uttered by six male speakers (MHO, MHT, MMY, MSH, MTK, and MYI) and four female speakers (FKN, FKS, FTK, and FYM), and a speech database containing the same sentences as those in the ATR Japanese speech database but uttered by a different female speaker (FTY) and a different male speaker (MMI). We chose four of these males (MHO, MMY, MSH, and MYI) and four of these females (FKN, FKS, FYM, and FTY) as training speakers for the average voice model and used the other three males (MHT, MTK, and MMI) and the other female (FTK) as target speakers of the speaker adaptation. These speech databases consist of high-quality, clean speech data collected under controlled recording studio conditions.

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 [48], [63] including the zeroth coefficient, $\log F_0$, and their delta and delta-delta coefficients. We used 5-state left-to-right context-dependent HSMMs without skip paths. Each state had a single Gaussian pdf with a diagonal covariance matrix as the state output pdf and had a single Gaussian pdf with a scalar variance as the state duration pdf. The basic structure of the speech synthesis system is the same as that in [20]. In the modeling of the synthesis units, we used 42 phonemes, including silence and pause, and took into account the phonetic and linguistic contexts [19]. Since 450 sentences from each training speaker were used for the training of the average voice models, 1800 sentences were used for training the gender-dependent average voice models and 3600 sentences were used for training the gender-independent average voice model. We first trained speaker-independent monophone HSMMs using manually annotated phonetic labels. These were converted into context-dependent HSMMs, and the model parameters were reestimated again. Then, shared-decision-tree-based context clustering [19] (using a minimum description length (MDL) criterion [64]) was applied to the HSMMs, and the model parameters of the HSMMs at each leaf node of the decision trees were tied. Note that decision trees were separately constructed for each state of Mel-cepstrum, $\log F_0$, and duration parts. We then reestimated the clustered HSMMs using SAT with piecewise linear regression functions and built the average voice models. [20]. We then adapted the average voice model to the target speaker. In the speaker adaptation and speaker-adaptive training, the estimation of multiple transforms was 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 transforms and hyperparameters of the MAP estimation, were manually and appropriately adjusted using objective measures explained in the next subsection. The thresholds to control the number of transforms were sensitive to the performance. We discuss it in Section IV-D. On the other hand, the hyperparameters of the MAP estimation were very less sensitive to the performance. The linear transforms $\boldsymbol{W}$ for the output pdfs in the linear regression algorithms were diagonal triblocks [54] corresponding to the static, delta, and delta-delta coefficients as follows:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_S & 0 & 0 \\ 0 & \boldsymbol{W}_\Delta & 0 \\ 0 & 0 & \boldsymbol{W}_{\Delta^2} \end{pmatrix} \tag{34}$$

where 0 is the square zero matrix, and $\boldsymbol{W}_S, \boldsymbol{W}_\Delta, \boldsymbol{W}_{\Delta^2}$ are the full square matrices for transformation of the static, delta, and delta-delta coefficients, respectively.

B. Evaluation of Initial Models

Parameter estimation for the speaker adaptation algorithms is an iterative procedure and its likelihood function has a lot of local maxima. The performance of the speaker adaptation therefore depends on the initial model. To investigate the effect of the initial model on the speaker adaptation, we first compared the adaptation of speaker-dependent (SD), gender-independent (GI), and gender-dependent (GD) models to the target speakers.

[3]http://www.atr-p.com/sdb.html

The eight speaker-dependent models were trained from 450 sentences of each training speaker by using the HSMM-based speaker-dependent method described in [29]. Here the adaptation method of each model was MLLR.

The objective measures we calculated were Euclidean distances of the acoustic features used: the target speakers' average Mel-cepstral distance and root-mean-square-error (RMSE) of $\log F_0$. After 50 sentences were used for adaptation, 50 test sentences included in neither the training nor the adaptation data were used for the evaluation. For the calculation of the average Mel-cepstral distance and the RMSE of $\log F_0$, the state duration of each HSMM model was adjusted after Viterbi alignment with the target speakers' actual utterance. Silence and pause regions were eliminated from the Mel-cepstral distance calculation. Since F_0 is not observed in the unvoiced region, the RMSE of $\log F_0$ was calculated in the region where both the generated and the actual F_0 were voiced.

The average measures between the actual and synthetic speech of the four target speakers are listed in Table II. From the results of the eight speaker-dependent models, we list only the results of the best and worst models (The label "GD+GD" in the table is explained later in this subsection). From the data listed in this table, we can first see that the use of the speaker-dependent model as an initial model is a risky and unsmart strategy, since the performance differs widely in individual speaker-dependent models. To choose the best speaker-dependent model, we then need to calculate some measures between target speakers and training speakers. We can also see that the best speaker-dependent model does not outperform the gender-dependent models. Comparing the gender-independent and gender-dependent models, we see that the RMSE of $\log F_0$ for gender-independent model is slightly worse than that for gender-dependent model, whereas the Mel-cepstral distances for gender-dependent and gender-independent models are about the same. Thus, the use of gender-dependent models seems to generally be a reasonable choice.

The voice and prosodic characteristics of some male speakers, however, are closer to those of a female or gender-independent average voice model, and the voice and prosodic characteristics of some female speakers are closer to those of a male or gender-independent average voice model. For the male speaker MMI, for example, the gender-independent model works better than the gender-dependent model with regard to the Mel-cepstral distance and the RMSE of $\log F_0$. We therefore investigated the effect on the simultaneous use of gender-dependent models to perform soft decisions of the models. For this purpose, we used a multiple linear regression approach [42] combining both the gender-dependent models

$$\boldsymbol{\mu}_i = \boldsymbol{\zeta}^{\mathrm{MALE}}\boldsymbol{\mu}_i^{\mathrm{MALE}} + \boldsymbol{\zeta}^{\mathrm{FEMALE}}\boldsymbol{\mu}_i^{\mathrm{FEMALE}} + \boldsymbol{\epsilon} \qquad (35)$$

where $\boldsymbol{\mu}_i^{\mathrm{MALE}}$ and $\boldsymbol{\mu}_i^{\mathrm{FEMALE}}$ are mean vectors for state i in the male and female average voice models. The speaker-adaptive training of the average voice models also used this multiple linear regression. The results of this multiple linear regression of the gender-dependent models are listed in the rows labeled "GD+GD" in the five parts of Table II, where one finds a slight improvement over the use of a single gender-dependent model only in the Mel-cepstral distance of the male speaker MMI. Since this multiple-regression approach requires about twice as many parameters to be estimated from the limited amount of adaptation data, it seems to suffer from decrease of the accuracy of the estimation.

TABLE II
RESULTS OF EVALUATION OF THE INITIAL MODELS. (THE ADAPTATION ALGORITHM WAS MLLR, AND THERE WERE 50 ADAPTATION SENTENCES.) (a)TARGET SPEAKER MHT. (b) TARGET SPEAKER MTK. (c) TARGET SPEAKER MMI. (d) TARGET SPEAKER FTK. (e) AVERAGE

(a)

Initial Model	Mel-cepstral distance (dB)	RMSE of $\log F_0$ (cent)
Worst SD	5.1	370
Best SD	4.8	300
GI	4.6	300
GD	4.7	290
GD+GD	4.7	290

(b)

Initial Model	Mel-cepstral distance (dB)	RMSE of $\log F_0$ (cent)
Worst SD	5.5	310
Best SD	5.1	260
GI	4.9	270
GD	4.9	260
GD+GD	4.9	260

(c)

Initial Model	Mel-cepstral distance (dB)	RMSE of $\log F_0$ (cent)
Worst SD	5.3	320
Best SD	5.1	280
GI	4.9	260
GD	5.0	270
GD+GD	4.9	270

(d)

Initial Model	Mel-cepstral distance (dB)	RMSE of $\log F_0$ (cent)
Worst SD	6.0	280
Best SD	5.7	240
GI	5.5	240
GD	5.5	230
GD+GD	5.6	240

(e)

Initial Model	Mel-cepstral distance (dB)	RMSE of $\log F_0$ (cent)
Worst SD	5.5	320
Best SD	5.2	270
GI	5.0	270
GD	5.0	260
GD+GD	5.0	270

To confirm the effect of the initial models from the viewpoint of perceptual differences, we conducted an ABX comparison test. The single use of the gender-independent and gender-dependent models and the simultaneous use of the gender-dependent models were investigated. We compared similarity of the synthetic speech using the models adapted from those initial models. In the ABX test, A and B were a pair of synthetic speech samples generated from the two models randomly chosen from

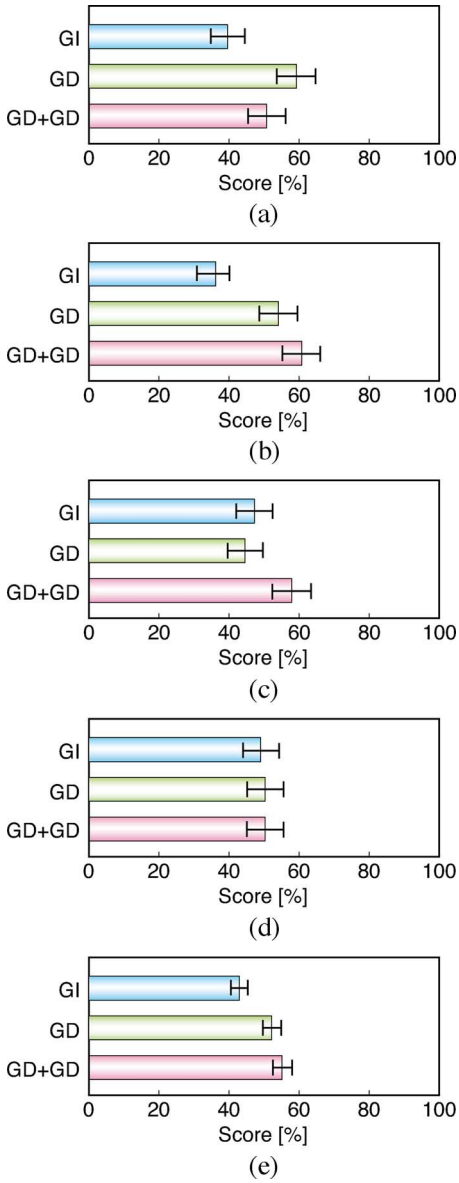


Fig. 7. Subjective evaluation of the initial models using ABX tests. (The adaptation algorithm was MLLR, and there were 50 adaptation sentences.) (a) Target speaker MHT. (b) Target speaker MTK. (c) Target speaker MMI. (d) Target speaker FTK. (e) Average.

the above combinations of the initial models, and X was the reference speech. The reference speech was synthesized by a Mel-cepstral vocoder since the quality of the speech by "copy-synthesis" can be considered as an upper bound on the performance of synthetic speech using parameters generated from HMMs and the same vocoder. Seven subjects were presented synthetic speech in the order of A, B, X or B, A, X, and were asked to whether the first or second speech sample was more similar to X (In an XAB test, the order becomes X, A, B or X, B, A). For each subject, four test sentences were randomly chosen from the same set of test sentences.

The average preference scores obtained in the ABX test are shown in Fig. 7 along with their 95% confidence intervals. From this figure we can see that these subjective evaluation results

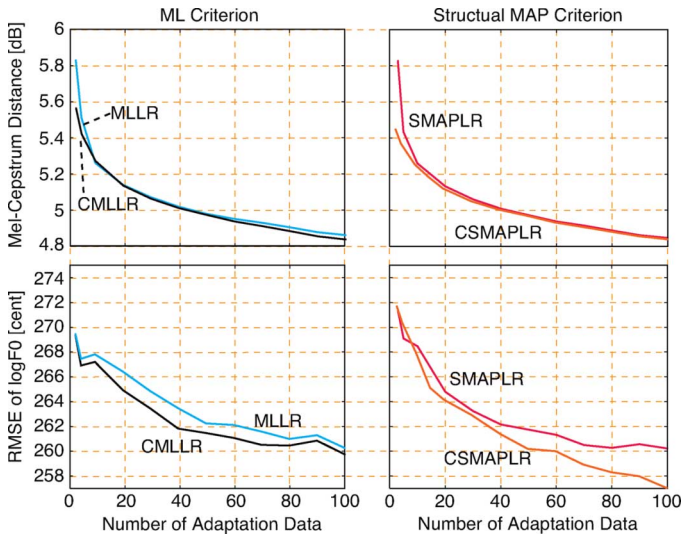


Fig. 8. Objective evaluation of transform functions in linear regression algorithms. Upper: average Mel-cepstral distance [dB], lower: RMSE of $\log F_0$ [cent].

are consistent with the above objective results—the gender-dependent models are better initial models than the gender-independent model is, and the simultaneous use of the gender-dependent models does not provide a significant improvement for these speakers. We therefore concluded that the gender-dependent models are a reasonable choice for the initial models for speaker adaptation.

C. Evaluation of Transform Functions and Estimation Criteria

We then compared the transform functions in the linear regression algorithms. To assess the effect on the covariance transform in the ML criterion and the SMAP criterion, we measured CMLLR and CSMAPLR against MLLR and SMAPLR. The transform 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. The ML criterion is used in MLLR and CMLLR, whereas the SMAP criterion is used in SMAPLR and CSMAPLR. From 5 to 100 adaptation sentences were used, and the target speakers and other experimental conditions were the same as those described in Section IV-B. The objective measures of the distance between the actual and synthetic speech of the target speakers are shown in Fig. 8, from which one sees that the mean and covariance transform in CMLLR and CSMAPLR adaptation generally produces a better RMSE of $\log F_0$ than the mean transform in MLLR and SMAPLR adaptation does but decreases the average Mel-cepstral distance only slightly.

We next compared the estimation criteria in the linear regression algorithms. To assess the effect on the estimation criteria in each transform function, we measured SMAPLR and CSMAPLR against MLLR and CMLLR. The experimental conditions were the same as those in the above comparison of transform functions. The objective measures of the distance between the actual and synthetic speech of the target speakers are shown in Fig. 9, where one can see that the adaptation algorithms using

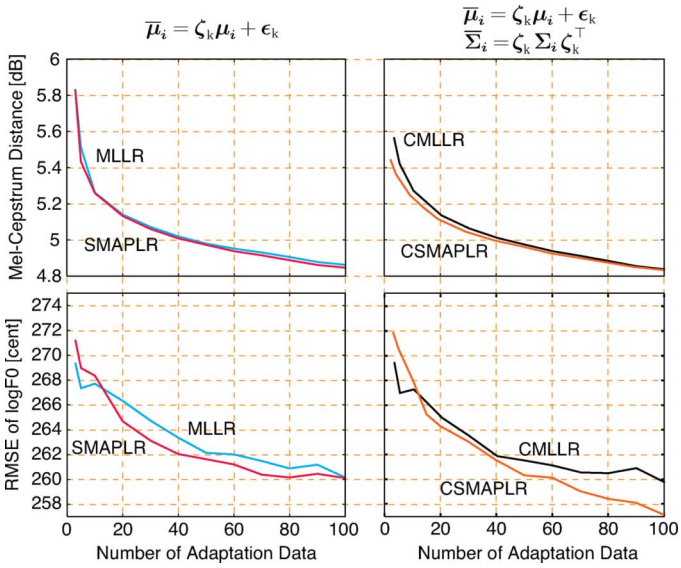


Fig. 9. Objective evaluation of estimation criteria in linear regression algorithms. Upper: average Mel-cepstral distance [dB], lower: RMSE of $\log F_0$ [cent].

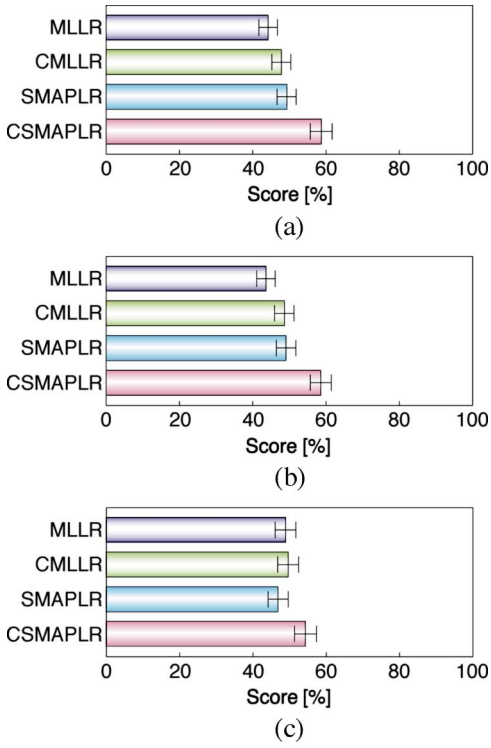


Fig. 10. Subjective evaluation of the similarity of synthetic speech generated from models adapted using several linear regression algorithms. (There were 50 adaptation sentences.) (a) Target speaker MHT. (b) Target speaker MTK. (c) Target speaker MMI.

the SMAP criterion also produce a better RMSE of $\log F_0$ than those using the ML criterion do. These results indicate that the CSMAPLR adaptation, which uses both the covariance transform and the SMAP criterion, would make the prosody of synthetic speech more similar to that of actual speech than the conventional adaptation techniques do.

To confirm the improvements of the CSMAPLR adaptation algorithm, we used the XAB comparison test to evaluate the similarity of the synthetic speech generated from the adapted models. The target speakers were MHT, MTK, and MMI. We excluded the target speaker FTK from this subjective evaluation, because the objective measures of the linear regression algorithms for the speaker show a similar tendency to the speaker MTK. Fifty adaptation sentences were used for this evaluation. Seven subjects were presented first with the reference speech sample and then with a pair (in random order) of the synthetic speech samples generated from two adapted models chosen from MLLR, CMLLR, SMAPLR, and CSMAPLR. The subjects were then asked which sample was closer to the reference speech. For each subject, eight test sentences were randomly chosen from fifty test sentences contained in neither the training nor the adaptation data.

The preference scores and 95% confidence intervals are shown in Fig. 10. The results confirm that although either the covariance transform only (CMLLR) or the SMAP criterion only (SMAPLR) does not produce synthetic speech significantly more similar to the target speaker's speech than the synthetic speech produced by the conventional MLLR, the CSMAPLR adaptation using both the covariance transform and the SMAP criterion does.

D. Evaluation of Sensitivity

Since the above linear regression algorithms use multiple transforms, we generally need some thresholds or parameters for controlling the number of the transforms. In the previous experiments, the appropriate numbers of the transforms were manually determined as mentioned earlier. In a practical sense, however, the adaptation performance of those piecewise linear regression algorithms depends on their threshold sensitivities. The adaptation performance is especially dependent on these threshold sensitivities when the target speakers are unknown/undefined/unfixed. We therefore investigated the threshold sensitivity of each algorithm and compared it with those of the other algorithms. The target speakers were MHT, MTK, and MMI, and there were 50 adaptation sentences from each of these speakers. Experimental conditions other than the thresholds were the same as those in the experiments described in Section IV-C.

We gradually increased the number of the transforms for each algorithm and calculated the objective measures. As the number of transforms increases, we can perform detailed piecewise linear regression and utilize more linguistic information of the tree structures. When the number of transforms increases more than necessary, however, the amount of adaptation data used for estimating a single transform relatively decreases and over-fitting to the adaptation data occurs. This results in degradation of the quality of synthetic speech. In addition, the rank-deficient problem occurs because to estimate multiple transforms we need to calculate several inverse matrices. Although we could use the generalized inverses matrices with singular value decomposition, this would decrease the accuracy of the estimation. In MLLR and SMAPLR this rank-deficient problem occurs when the number of distributions that share a single transform is less than the number of dimensions of the

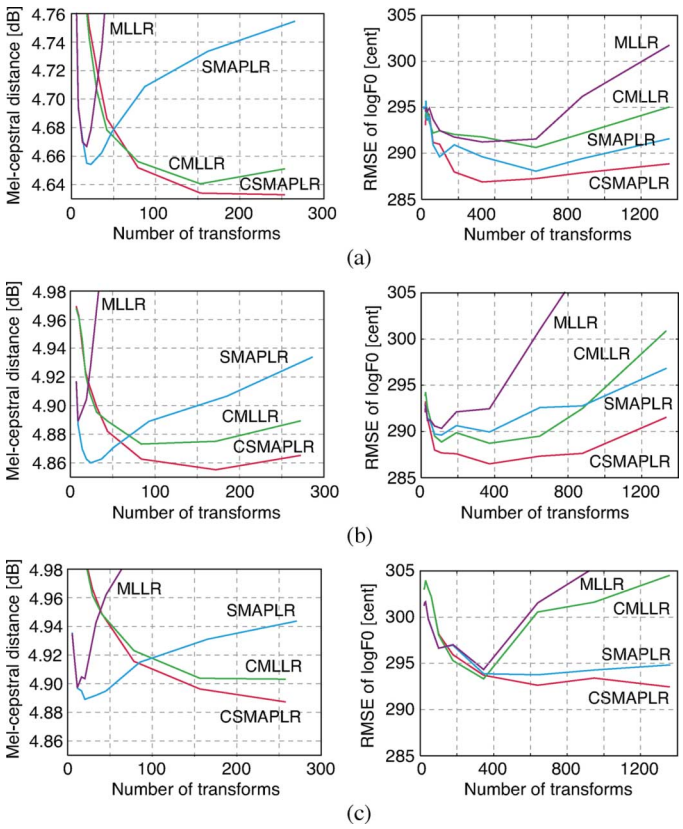


Fig. 11. Objective evaluation of sensitivity of linear regression algorithms. (There were 50 adaptation sentences.) Left: average Mel-cepstral distance [dB], right: RMSE of $\log F_0$ [cent]. (a) Target speaker MHT. (b) Target speaker MTK. (c) Target speaker MMI.

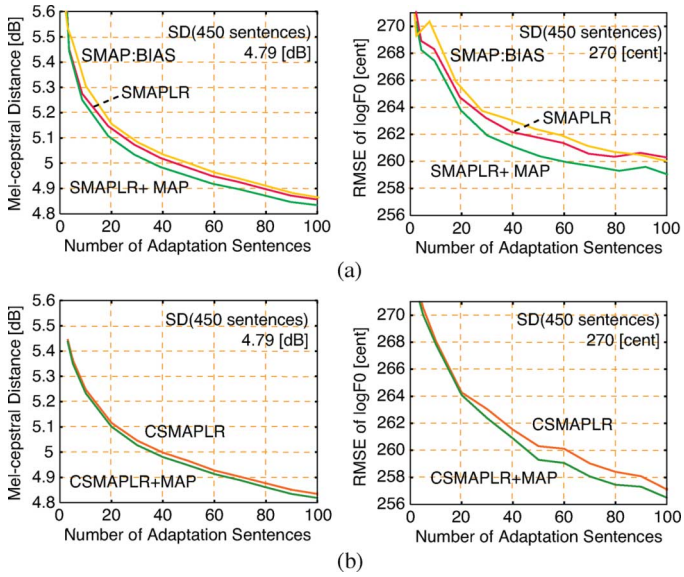


Fig. 12. Objective evaluation of linear regression algorithms combined with MAP adaptation. Left: average Mel-cepstral distance [dB], right: RMSE of $\log F_0$ [cent]. (a) SMAPLR+MAP. (b) CSMAPLR+MAP.

feature vector. On the other hand, in CMLLR and CSMAPLR it occurs when the number of the observations vectors to be used for the estimation of a single transform is less than the number of dimensions of the feature vector. The rank-deficient problem would therefore occur more easily in MLLR and SMAPLR than in CMLLR and CSMAPLR because increasing the number of transforms in MLLR and SMAPLR is directly linked to the rank-deficient problem.

The objective evaluation results for the sensitivity of MLLR, CMLLR, SMAPLR, and CSMAPLR are shown in Fig. 11, where the horizontal axis represents the number of transforms, which were determined via the thresholds. From Fig. 11 we can clearly see that the MLLR adaptation algorithm is the one most sensitive to the change in the thresholds for controlling the number of transforms. Especially, it is sensitive to the change of the thresholds for Mel-cepstral coefficients. Comparing CMLLR with MLLR, we see that the CMLLR adaptation is less sensitive to the change in the thresholds. One of the reasons would be due to the above rank-deficient problem. Comparing SMAPLR with MLLR or CSMAPLR with CMLLR, we notice that the robust SMAP criterion can alleviate the over-fitting problem and that the thresholds used for SMAPLR and CSMAPLR become less sensitive than those used for MLLR and CMLLR. As a consequence, we can confirm that the CSMAPLR adaptation is the one least sensitive to the change of the thresholds for controlling the number of the transforms. It is thus expected to work stably and robustly even for new features or new speakers.

E. Evaluation of Combined Algorithms

As mentioned previously, in the speaker adaptation using linear regression, there is a rough assumption that the target speaker model would be expressed by the piecewise linear regression of mean vectors of the average voice model. We therefore investigate the effect of the algorithms combining linear regression and MAP adaptation [41]. By performing MLLR or SMAPLR adaptation followed by MAP adaptation, it would be possible to appropriately modify the mean vectors of the distributions having relatively sufficient amount of speech samples. By performing CMLLR or CSMAPLR adaptation followed by MAP adaptation, it would be possible to perform sphering (whitening) of diagonal covariance as well as the modification of the mean vectors.

We compared the synthetic speech obtained before and after applying the combined algorithm to models created by SMAPLR or CSMAPLR adaptation. For the combination with SMAPLR adaptation, we updated only the mean vectors in order to see the effect of their modification. For the combination with CSMAPLR adaptation, we updated both the mean vectors and the covariance matrices. The target speakers were MHT, MTK, MMI, and FTK, and the experimental conditions were the same as those described Section IV-C. The objective measures between the actual and synthetic speech obtained with and without the additional MAP adaptation of the target speakers are shown in Fig. 12. The results of the combination with SMAPLR adaptation (SMAPLR+MAP) are shown in part (a), and the results of the combination with CSMAPLR adaptation (CSMAPLR+MAP) are shown in part (b). The objective measures for speaker dependent (SD) models trained with 450 sentences and bias adaptation using the SMAP criterion (SMAP:BIAS) are also shown for reference. The improvement

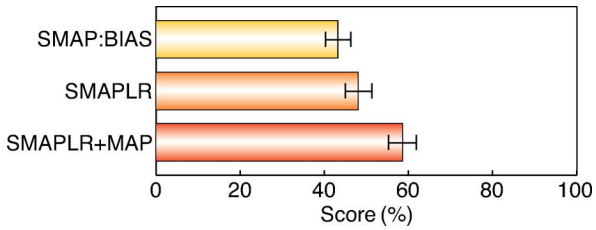


Fig. 13. Subjective evaluation of SMAP-based linear regression algorithms and an algorithm combining MAP adaptation with SMAP-based linear regression. (There were 50 adaptation sentences.)

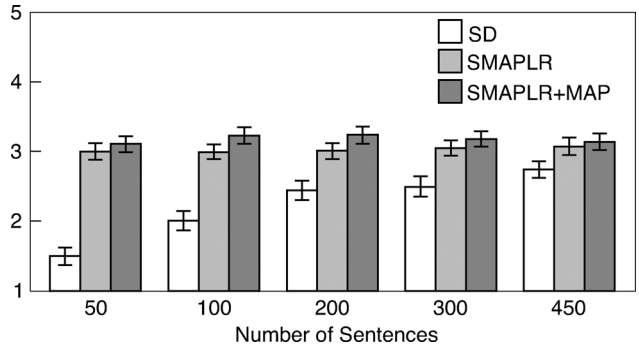


Fig. 14. Subjective evaluation of models adapted using SMAPLR or using SMAPLR combined with MAP adaptation. Similarity was rated on a 5-point scale: 5 for very similar, 4 for similar, 3 for slightly similar, 2 for dissimilar, and 1 for very dissimilar.

due to the combined algorithm with MAP adaptation is evident when there are at least 20 sentences, and the additional MAP adaptation seems to improve the RMSE of $\log F_0$ more than it does the Mel-cepstrum distance. We also see that the model adapted using 100 sentences gives results comparable to the speaker-dependent model in terms of the Mel-cepstral distance and a little better than the speaker-dependent model in terms of the RMSE of $\log F_0$. This tendency is similar to that in the results reported in [20].

Although the above evaluations show that the CSMAPLR adaptation worked better than SMAPLR adaptation, we used the XAB test to compare similarity of the synthetic speech generated from the models adapted using SMAP-based bias adaptation, SMAPLR adaptation, and a combination of MAP adaptation and the SMAPLR adaptation so that we could assess the effects of the following transforms of the mean vectors

$$\text{SMAP : BIAS} \quad \boldsymbol{\mu}^{\text{BIAS}} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{36}$$

$$\text{SMAPLR} \quad \boldsymbol{\mu}^{\text{LINEAR}} = \boldsymbol{\zeta}\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{37}$$

$$\text{SMAPLR+MAP} \quad \boldsymbol{\mu}^{\text{MAP}} = (1-\xi)\boldsymbol{\mu}^{\text{ML}} + \xi\boldsymbol{\mu}^{\text{LINEAR}} \tag{38}$$

where $\boldsymbol{\mu}^{\text{ML}}$ is an ML estimator of the mean vector calculated from the adaptation data and ξ is a hyperparameter for determining the interpolation ratio between the ML estimator and the mean vector transformed by the linear regression. The target speakers were the same four speakers. Seven subjects were presented first with reference speech and then with a pair of synthetic speech (in random order) samples generated from the models. The subjects were then asked which synthetic speech was more similar to the reference speech. For each subject, five test sentences were randomly chosen from a set of test sentences contained in neither the training nor the adaptation data.

The preference scores (with 95% confidence intervals) of each speaker adaptation algorithm are shown in Fig. 13, where we can see that the subjects can distinguish between synthetic speech obtained using the linear regression algorithms and synthetic speech obtained using the combined algorithm. This implies that the target acoustic features cannot be perfectly reproduced by only the (piecewise) linear transforms of the average voice model. Although we need additional forward-backward calculations after the linear transforms, it is worth using the combined algorithm to reduce the gap between the ML estimator and linearly transformed parameters.

For evaluating the effect of the amount of adaptation data used in the above combined algorithm, we conducted a comparison category rating (CCR) test. For reference, we also evaluated SD models trained on each of the amounts of adaptation data that were used to transform the models using SMAPLR adaptation and the algorithm combining SMAPLR and MAP adaptation. The numbers of adaptation sentences were 50, 100, 200, 300, and 450. The target speakers were the same four speakers. Seven subjects were presented first with reference speech and then, in random order, with synthetic speech samples generated from the models. The subjects were then asked to rate the similarity of the synthetic speech to the reference speech. The rating was done using a 5 point scale: 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 contained in neither the training nor the adaptation data.

The average scores obtained in this CCR test are shown in Fig. 14 along with 95% confidence intervals. In this figure, contrary to our expectations that the combined algorithm gradually improves the quality of synthetic speech as the amount of the adaptation sentences increases, we can see significant differences between the linear regression algorithm and the combined algorithm only for 100 and 200 adaptation sentences. Even for other amount of adaptation data, however, the scores for the combined algorithm are at least comparable to those for the linear regression algorithm. Thus, we can say that although the combined algorithm does not improve the similarity of synthetic speech greatly, it does improve it. We used the SMAPLR adaptation in these evaluations, but an algorithm combining MAP adaptation with CSMAPLR adaptation would have a similar effect because CSMAPLR adaptation is also a piecewise linear regression algorithm.

It is obvious that with a limited amount of adaptation sentences the average similarity scores for synthetic speech obtained using the combined algorithm and even the linear regression algorithm are significantly better than those for the synthetic speech obtained using the speaker-dependent approach. What is very interesting about this result is that the average scores for synthetic speech obtained using the speaker adaptation algorithms are still significantly better than those for the speaker-dependent approach even for all the larger amounts of adaptation sentences. This seems paradox because 1) the purpose of the combined algorithm is to reduce the gap between the ML estimators of the mean vectors calculated from

the adaptation data and linearly transformed mean vectors and then 2) the speaker-dependent models are also built using the same number of sentences with an ML criterion. As shown in Fig. 13, the model using the combined algorithm is better than linearly transformed model since the linearly transformed mean vectors are modified by the ML estimators of the mean vectors. However, the speaker-dependent model, which holds *other* ML estimators of the mean vectors, is worse than one using the combined algorithm in Fig. 14. We should therefore consider the possibility that these results are due to some other factors in the HMM training process. One would be relation between the amount of training data for the average voice model and decision-tree-based clustering of Gaussian distributions of the HMMs. To cope with problems of data sparsity and unseen context-dependent HMMs, we use the MDL criterion and build decision trees for clustering of distributions. The decision trees for the average voice model, which can easily utilize a lot of speech data, thereby generally become larger and more precise than those for the speaker-dependent model. The decision-tree size differences caused by the amount of the training data would affect the above results.

F. Evaluation of the Amount of Training Data for the Average Voice Model

We first conducted an objective evaluation for confirming our hypothesis on the effect of the decision-tree size differences. The easiest way is to eliminate the influence of the different decision-tree size and topology and compare the performance. We thus constructed decision trees having the same structure/topology and size common to all the training speakers for the average voice model and target speakers using the shared-tree-based clustering algorithm [34]. The target speaker used was a male speaker MTK. The training speakers used were five male speakers (MHO, MHT, MMY, MSH, and MYI). The gender-dependent average voice model was trained using 450 sentences for each training speaker, 2250 sentences in total. The average voice model was then adapted to the target speaker using 5 to 450 sentences. The SD model having the common decision trees was also built using 450 sentences of the target speaker. Other experimental conditions were the same as those described Section IV-C.

The objective measures between the actual and synthetic speech obtained from models using several adaptation algorithms and the SD models are shown in Fig. 15. The Mel-cepstral distance is shown in part (a) and the RMSE of $\log F_0$ is shown in part (b). We can see that these results using the common decision trees show obviously different trends. The measures of the combined algorithm (SMAPLR+MAP) asymptotically come close to those of the SD models as the amount of adaptation sentences available increases. This is consistent with the nature of MAP estimation [24]. The important thing to remember is, however, that this situation does not happen in reality because the average voice model and SD model may always have individual tree structures/topologies and sizes, resulting in mismatch in Fig. 14. Moreover the tree

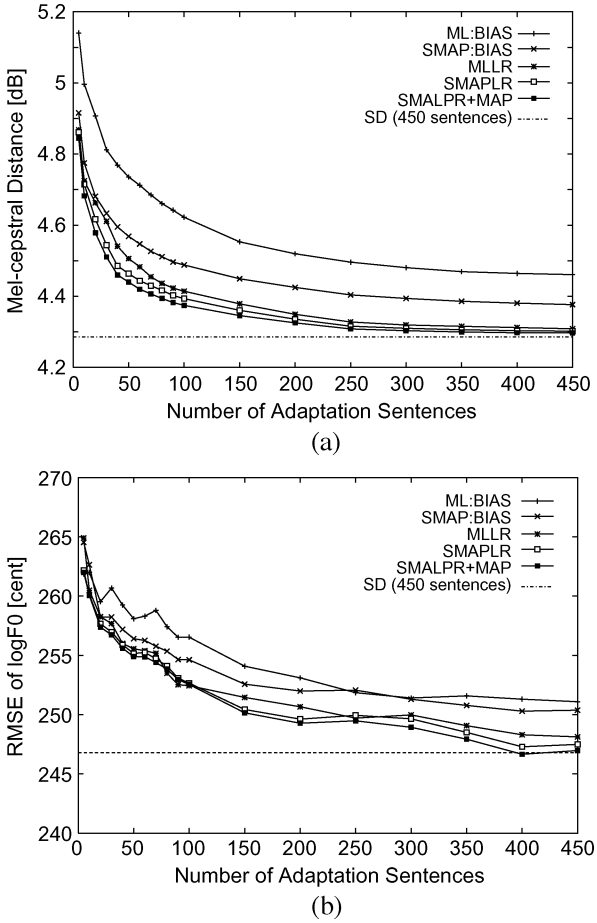


Fig. 15. Objective evaluation of several speaker adaptation algorithms. The common decision trees are used in both the average voice model and speaker-dependent model to eliminate the influence of the different decision-tree size and topology. (a) Average Mel-cepstral distance [dB]. (b) RMSE of $\log F_0$ [cent].

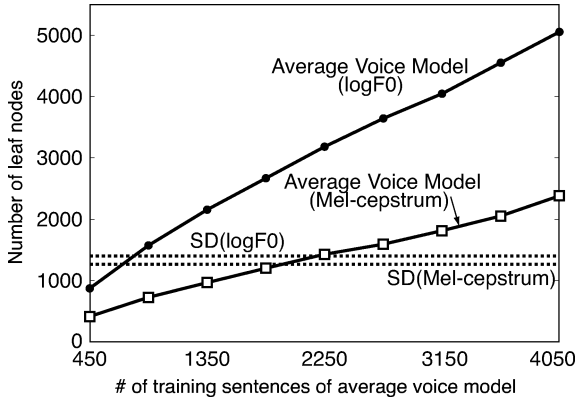


Fig. 16. Number of leaf nodes of the decision trees for the average voice models versus the number of the training sentences for the average voice models.

structure and size for the average voice model may vary with the amount of the training data available.

Therefore, we investigated the effect of the amount of training data for the average voice models and analyzed how the tree structure and size for the average voice model are associated with the performance of the adaptation. Although the use of the gender-dependent average voice models are, as discussed above, more appropriate than the gender-independent

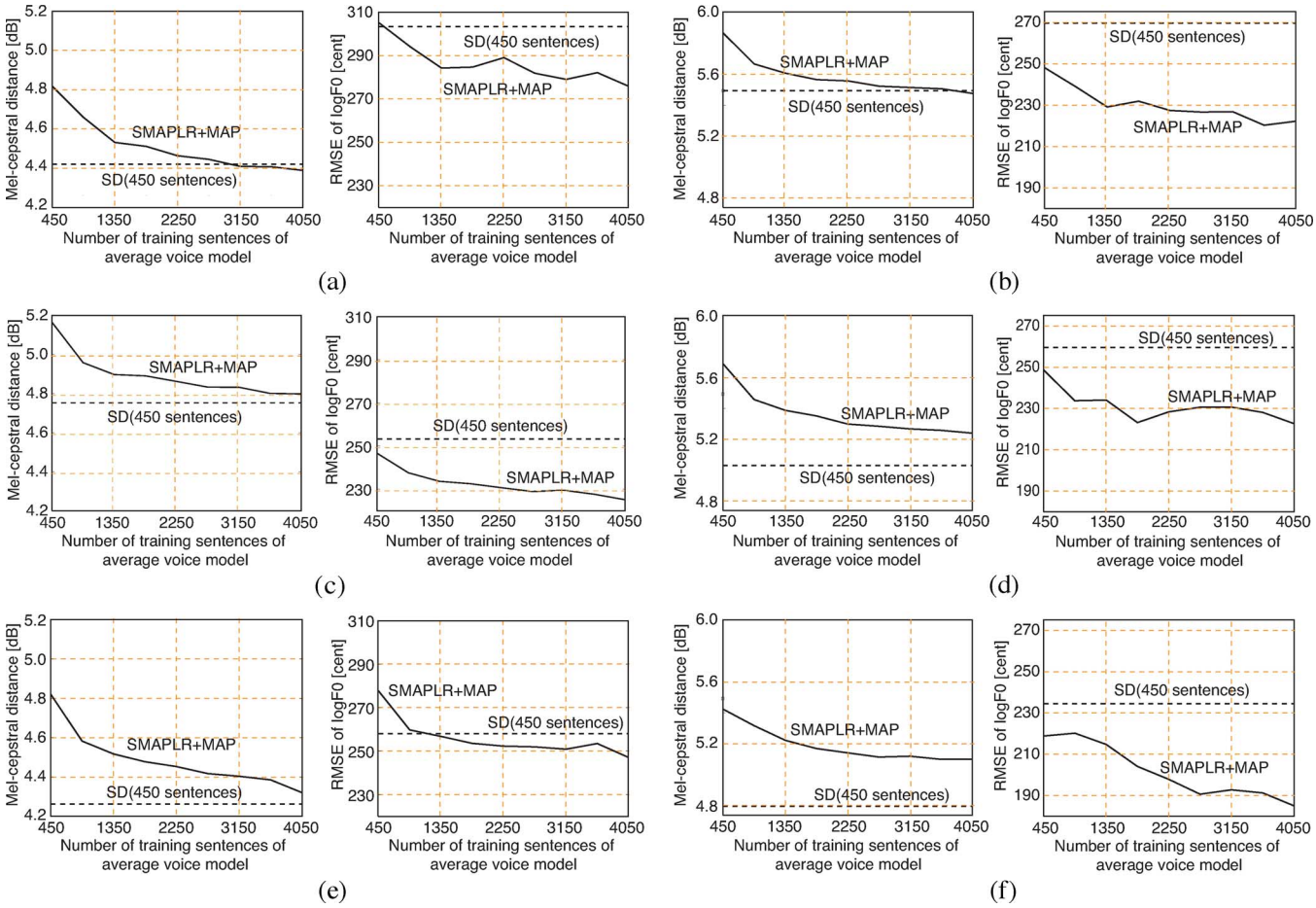


Fig. 17. Objective evaluation of the amount of training data for the average voice models. (The adaptation data was 450 sentences, and the adaptation algorithm was the one combining SMAPLR with MAP.) Left: average Mel-cepstral distance [dB], right: RMSE of $\log F_0$ [cent]. (a) Target speaker MHT. (b) Target speaker FKS. (c) Target speaker MSH. (d) Target speaker FTK. (e) Target speaker MTK. (f) Target speaker FTY.

model average voice model, we used the gender-independent one in these experiments because it makes it much easier to see the effects of the amount of training data. From the speakers included in the speech database described above, we chose as the target speakers one male and one female speaker and trained a gender-independent average voice model using the speech data obtained from the rest of the five speakers (excluding MMI) and four female speakers. By changing the combination of the target speakers and training speakers, we evaluated total of six speakers (MHT, MSH, MTK, FKS, FTK, and FTY) as the target speakers. The amount of the training data used for the gender-independent average voice models was ranged from 450 to 4050 sentences. The amount of the speech data used from each training speaker was ranged from 50 to 450 sentences in increments of 50 sentences. For reference, the speaker-dependent models trained from the 450 sentences were also evaluated at the same time. The same sentence set was used for the average voice models and the speaker-dependent models. The sentence set consists of several predefined subsets [65]. To reduce the bias of contextual information between the subsets included in the training data for the average voice model, we worked out the combination of the subset of training data from each training speaker so that we could reduce the overlap between the subsets as much as possible. Thus, even the minimum amount of the training data (450 sentences) contains the same sentences set as that of the speaker-dependent models. We used 450 sentences from the target speaker as adaptation data in order to assess the effect of the amount of the training data and the number of leaf nodes of the decision trees as accurately as possible. The adaptation algorithm we used was the one combining SMAPLR adaptation and MAP adaptation. Although we should have used CSMAPLR adaptation rather than SMAPLR adaptation, we used SMAPLR adaptation method in order to reduce the computational time needed for the relatively large amount of adaptation data. Even if we used the CSMAPLR adaptation in these experiments, it would not change the results very much because the MAP adaptation using the relatively generous amount of adaptation data is applied to the model transformed by SMALR adaptation. The experimental conditions were the same as those in the experiment described in Section IV-C.

The average number of leaf nodes of decision trees constructed for the average voice models is shown in Fig. 16 as a function of the number of training sentences for the average voice models. In this figure, we also show the average number of leaf nodes of decision trees used for the SD models. In construction of all the decision trees, the MDL criterion was used for preventing over-fitting and determining an appropriate

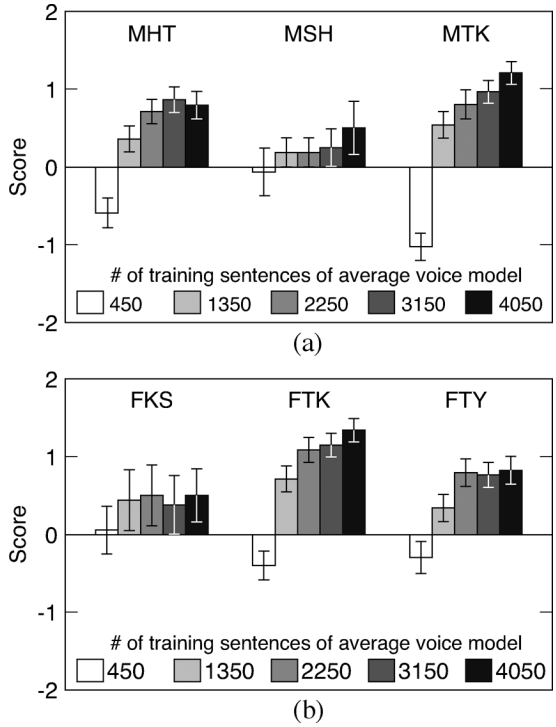


Fig. 18. Subjective evaluation of the amount of the training data for the average voice models. The evaluation method was the CCR test, and there were adaptation sentences. The adaptation algorithm was SMAPLR+MAP, and naturalness was rated on a 5-point scale: $+2$ for much more natural than the SD models, 0 for almost the same as the SD models, and -2 for much less natural than the SD models. (a) Male speakers. (b) Female speakers.

number of the leaf nodes. The MDL criterion can be explained simply; it ensures that the increase in likelihood after each split of nodes of the decision tress is above a threshold dictated by the feature dimensionality and the number of observation frames in the training set [64]. From this figure we can see that the number of leaf nodes increases linearly as the amount of training data for the average voice model increases. Especially, the decision trees for $\log F_0$ are much larger than those of the SD models.

We calculated the objective measures of the difference between actual and synthetic speech, and for reference we also calculated the objective measures for the SD models. Fig. 17 shows the objective measures for each target speaker as a function of the number of the training sentences for the average voice models. From these results it can be seen that both the Mel-cepstral distance and RMSE of $\log F_0$ of synthetic speech generated from the adapted model become remarkably better as the amount of training data for the average voice models increases. Although it is ironic that increasing the amount of training data is more effective than other speaker-adaptation-improving methods we investigated in the work reported in this paper, this is simple and effective. We can see in Fig. 17 that when more than 1350 sentences are used for training the average voice models, for all the target speakers the RMSE of $\log F_0$ is better with the adapted models than it is with the SD models. When 4050 sentences are used for training the average voice models, the Mel-cepstral distance for the target speakers MHT and FKS is slightly better with the adapted models than it is with the SD models. Since the improvements in both Mel-cepstral distance and RMSE of $\log F_0$ that are due to the increase of the amount of the training data for the average voice models do not converge even at 4050 sentences, they would be enhanced by using more than 4050 training sentences.

We used the CCR test to evaluate the naturalness of the synthetic speech obtained using each average voice model and SD model. Seven subjects were presented first with synthetic speech generated from the SD model as a reference and then with a speech sample generated from the adapted models randomly chosen from the set of the models. The subjects were then asked to rate the naturalness of the synthetic speech relative to that of the reference speech. The rating was done using the following five-point scale: $+2$ for much more natural, $+1$ for more natural, 0 for almost the same, -1 for less natural, and -2 for much less natural. For each subject, eight test sentences were randomly chosen from 50 test sentences that were contained in neither the training nor the adaptation data.

The results of the CCR test are shown (with 95% confidence intervals) in Fig. 18, from which we can see that, for all the target speakers, when there are more than 1350 training sentences the average scores for naturalness of synthetic speech obtained using the adapted models are higher than those of the synthetic speech obtained using the SD models. This result is consistent with the fact that when there are more than 1350 training sentences, the RMSE of $\log F_0$ of the speech obtained using the adapted models is better results than that of the speech obtained using the SD models. Although the Mel-cepstral distance of the speech obtained using adapted models were not better than those of the speech obtained using the SD models, the differences between the Mel-cepstral distance for the speech obtained using the adapted models and that for the speech obtained using SD models were small. When there were 1350 training sentences, this distance was between 0.1 and 0.3 dB. Thus, these differences would not have affected these subjective evaluation results. In fact, we can see that there is a strong correlation between the average scores for the naturalness of synthetic speech generated from the adapted models and the number of leaf nodes of the decision trees constructed for $\log F_0$ in Fig. 16. Since the number of leaf nodes of the decision tree increases linearly with the amount of training data for the average voice models, we can also say that the naturalness of synthetic speech generated from the adapted models is closely correlated with the amount of the training data for the average voice models. Using more training data is a very simple and straightforward but effective and reliable method for improving the quality of synthetic speech obtained using speaker adaptation methods.

Finally, we analyzed the influence of the tree-topology differences of the average voice model and SD models. In previous experiments, the MDL criterion is used for automatically determining an appropriate number of leaf nodes of the decision trees and thus the decision trees for the average voice model and SD models have different size. In order to eliminate the influence of the different decision-tree size, instead of the MDL criterion, we manually adjusted the number of leaf nodes of the decision trees for the SD models to that for the average voice model and compare the performance. As mentioned earlier, the MDL criterion specifies a threshold for the split of nodes of the

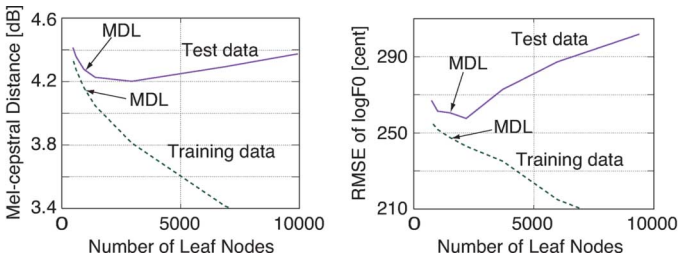


Fig. 19. Number of leaf nodes of the decision trees for the SD models versus the objective measures. (The training data used for the SD models was 450 sentences, and the speaker used was the male speaker MTK.) Left: average Mel-cepstral distance [dB], right: RMSE of $\log F_0$ [cent].

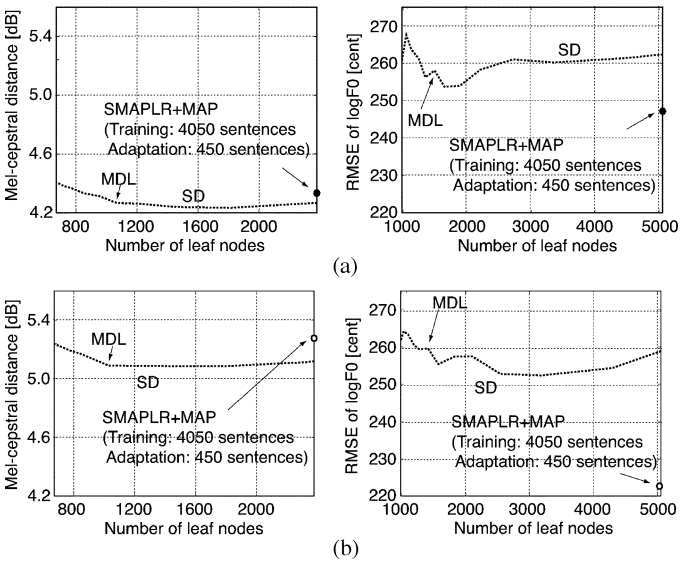


Fig. 20. Objective evaluation for the SD models with the same number of leaf nodes as those of the average voice model. (The training data used for the SD models was 450 sentences, and the speakers used was the male speaker MTK and the female speaker FTK.) Left: average Mel-cepstral distance [dB], right: RMSE of $\log F_0$ [cent]. (a) Target speaker MTK. (b) Target speaker FTK.

decision tress. We added a new positive weighting factor to the threshold and continued the split of the nodes by decreasing the weight. The speakers used were the male speaker MTK and the female speaker FTK. The amount of training data used for the SD models was 450 sentences. Fig. 19 shows the manually-adjusted numbers of leaf nodes for the male speaker MTK versus the objective measures. For reference, we calculated the objective measures for 50 sentences randomly selected from the training data as well as the 50 test sentences used in Fig. 17. From these figures, we can see that the number of leaf nodes determined by the MDL criterion is reasonably close to the optimal one for the test data, that is, the MDL criterion prevents over-fitting to the training data. Fig. 20 shows the objective evaluation results which focus on regions between the number of leaf nodes specified by the MDL criterion in the SD models and that specified by the MDL criterion in the average voice model. The average voice model was identical to one trained from 4050 sentences in Fig. 17. The results for MTK are shown in part (a) and those for FTK are shown in part (b). For reference, the adaptation results using 450 sentences were also shown in the figures. It can be seen that the RMSE of $\log F_0$ is still better with the adapted models than it is with even the SD models having the same number of leaf nodes as those of the average voice model. From these results, we can conclude the decision trees for the average voice model have a better topology for $\log F_0$ than those for the SD models.

V. Conclusion

In this paper, we have presented the results obtained using the subjective and objective tests to evaluate the effects of several factors and configuration choices encountered during training and model construction in speaker adaptation for HMM-based speech synthesis. We have also proposed the new robust and stable CSMAPLR adaptation algorithm whose derivation was based on the knowledge obtained in this analysis and by comparing the results of major linear-regression algorithms such as the MLLR, CMLLR, and SMAPLR adaptation algorithms. The findings obtained from our analysis results can be summarized as follows. Better and more stable adaptation performance can be obtained from a small amount of speech data by preparing gender-dependent average voice models from a large amount of training data as the initial model and adapting these models by using an algorithm combining CSMAPLR adaptation and MAP adaptation. Increasing the number of training sentences for the average voice model is a simple and remarkably effective method for improving the quality of synthetic speech obtained using speaker adaptation methods. It provides larger decision trees having a better topology. The CSMAPLR adaptation algorithm improves the RMSE of the $\log F_0$ of synthetic speech as well as the Mel-cepstral distance between actual and synthetic speech. It thereby improves the similarity of actual and synthetic speech. It can also reduce the threshold sensitivity for determining the number of multiple transforms. It is especially effective for the thresholds for Mel-cepstral coefficients. Algorithms combining linear regression with MAP adaptation are effective when more than 20 adaptation sentences are available. They would improve the similarity of actual and synthetic speech by appropriately modifying both the mean vectors and covariance matrices of the Gaussian distributions of the MSD-HSMMs. In addition the coherence between these subjective and objective tests reported in this paper itself are an interesting result.

This study on speaker adaptation will also have strong relevance to style adaptation since these factors also affect style adaptation. It would also be interesting to investigate other factors such as the number of the training speakers or the number of training speakers who are professional narrators. Our future work is to release these adaptation algorithms for speech synthesis to the public[4] and develop an unsupervised speaker adaptation algorithm for speech synthesis.

[4]Some of these techniques have already been released in an open-source software toolkit called HTS (from "H Triple S," an acronym for the "HMM-based speech synthesis system") [66]. The recently released HTS version 2.0 [67] includes the MLLR and CMLLR adaptation algorithms and algorithms combining them with MAP adaptation for MSD-HMMs. We also plan to integrate and release several additional adaptation algorithms including the SMAPLR and CSMAPLR algorithms for the MSD-HSMMs as a part of a new HTS version 2.1. Therefore, these results on the speaker adaptation techniques described in this paper would also be very beneficial to new users of the HTS toolkit who want to try speaker adaptation techniques using their speech data.

ACKNOWLEDGMENT

The authors would like to thank Prof. K. Tokuda and Dr. H. Zen of the Nagoya Institute of Technology and Prof. S. Renals, Dr. S. King, and Dr. K. Richmond of the University of Edinburgh for their valuable comments. They would also like to thank former students D. Nomura and T. Okawa for partly supporting our experiments.

REFERENCES

[1] A. Black and N. Cambpbell, "Optimising selection of units from speech database for concatenative synthesis," in *Proc. EUROSPEECH'95*, Sep. 1995, pp. 581–584.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP'96*, May 1996, pp. 373–376.

[3] R. Donovan and P. Woodland, "A hidden Markov-model-based trainable speech synthesizer," *Comput. Speech Lang.*, vol. 13, no. 3, pp. 223–241, 1999.

[4] A. Syrdal, C. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Storm, K. Lee, and M. Makashay, "Corpus-based techniques in the AT&T NEXTGEN synthesis system," in *Proc. ICSLP'00*, Oct. 2000, pp. 411–416.

[5] A. Black, "Unit selection and emotional speech," in *Proc. Eurospeech'03*, Sep. 2003, pp. 1649–1652.

[6] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP'95*, May 1995, pp. 660–663.

[7] K. Tokuda, T. Masuko, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from HMM using dynamic features," (in Japanese) *J. Acoust. Soc. Jpn.*, vol. 53, no. 3, pp. 192–200, Mar. 1997.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP'96*, May 1996, pp. 389–392.

[9] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "HMM-based speech synthesis using dynamic features," (in Japanese) *IEICE Trans.*, vol. J79-D-II, no. 12, pp. 2184–2190, Dec. 1996.

[10] 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*, Sep. 1999, pp. 2374–2350.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," (in Japanese) *IEICE Trans.*, vol. J83-D-II, no. 11, pp. 2099–2107, Nov. 2000.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP'00*, Jun. 2000, pp. 1315–1318.

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

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

[15] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[16] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.

[17] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP'97*, Apr. 1997, pp. 1611–1614.

[18] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP'01*, May 2001, pp. 805–808.

[19] 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.

[20] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[21] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *Proc. 3rd ESCA/COCOSDA Workshop Speech Synth.*, Nov. 1998, pp. 273–276.

[22] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system using MAP-VFS," (in Japanese) *IEICE Trans.*, vol. J83-D-II, no. 12, pp. 2509–2516, Dec. 2000.

[23] C. Leggetter and P. 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.

[24] 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.

[25] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," *Comput. Speech Lang.*, vol. 10, no. 2, pp. 117–132, 1995.

[26] J. Takahashi and S. Sagayama, "Vector-field-smoothed bayesian learning for fast and incremental speaker/telephone-channel adaptation," *Comput. Speech Lang.*, vol. 11, no. 2, pp. 127–146, 1997.

[27] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[28] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation of pitch and spectrum for HMM-based speech synthesis," (in Japanese) *IEICE Trans.*, vol. J85-D-II, no. 4, pp. 545–553, Apr. 2002.

[29] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[30] J. Ferguson, "Variable duration models for speech," in *Proc. Symp. Applicat. Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.

[31] M. Russell and R. Moore, "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. ICASSP'85*, Mar. 1985, pp. 5–8.

[32] S. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Lang.*, vol. 1, no. 1, pp. 29–45, 1986.

[33] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP'96*, Oct. 1996, pp. 1137–1140.

[34] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. Syst.*, vol. E86-D, no. 3, pp. 534–542, Mar. 2003.

[35] L. Qin, Z. Ling, Y. Wu, B. Zhang, and R. Wang, "HMM-based emotional speech synthesis using average emotion model," in *Proc. ISCSLP'06 (Springer LNAI Book)*, Dec. 2006, pp. 233–240.

[36] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Commun.*, vol. 48, no. 10, pp. 1227–1242, 2006.

[37] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

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

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

[40] O. Shiohan, T. Myrvoll, and C. Lee, "Structural maximum *a posteriori* linear regression for fast HMM adaptation," *Comput. Speech Lang.*, vol. 16, no. 3, pp. 5–24, 2002.

[41] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 294–300, Jul. 1996.

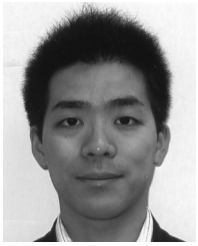
[42] J. Isogai, J. Yamagishi, and T. Kobayashi, "Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis," in *Proc. Eurospeech'05*, Sep. 2005, pp. 2597–2600.

[43] 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'06*, May 2006, pp. 77–80.

[44] 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'06*, Sep. 2006, pp. 2286–2289.

[45] 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'06*, Sep. 2006, pp. 1328–1331.

[46] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. ICASSP'07*, Apr. 2007, pp. 1233–1236.
[47] S. Imai, "Cepstral analysis synthesis on the Mel frequency scale," in *Proc. ICASSP'83*, Apr. 1983, pp. 93–96.
[48] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito, and S. Imai, "Spectral estimation of speech based on Mel-cepstral representation," (in Japanese) *IEICE Trans. Fundamentals*, vol. J74-A, no. 8, pp. 1240–1248, Aug. 1991.
[49] Speech Signal Processing Toolkit (SPTK) Version 3.1. 2007 [Online]. Available: http://www.sp-tk.sourceforge.net/
[50] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Inf. Syst.*, vol. E87-D, no. 12, pp. 2812–2820, Dec. 2004.
[51] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.
[52] J. Yamagishi, T. Kobayashi, S. Renals, S. King, H. Zen, T. Toda, and K. Tokuda, "Improved average-voice-based speech synthesis using gender-mixed modeling and a parameter generation algorithm considering GV," in *Proc. 6th ISCA Workshop Speech Synth.*, Aug. 2007, pp. 125–130.
[53] J. Yamagishi, T. Nose, H. Zen, T. Toda, K. Tokuda, S. King, and S. Renals, "A speaker-adaptive HMM-based speech synthesis for the Blizzard Challenge 2007," *IEEE Audio, Speech, Lang. Process.*, 2008, submitted for publication.
[54] M. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, no. 4, pp. 249–264, 1996.
[55] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.
[56] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Trans. Inf. Syst.*, vol. E91–D, no. 6, pp. 1764–1773, Jun. 2008.
[57] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 272–281, Mar. 1999.
[58] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. ICASSP'98*, May 1998, pp. 661–664.
[59] M. Rahim and B. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 19–30, Jan. 1996.
[60] M. Gales, "Multiple-cluster adaptive training schemes," in *Proc. ICASSP'01*, May 2001, pp. 361–364.
[61] A. Gupta and T. Varga, *Elliptically Contoured Models in Statistics*. Norwell, MA: Kluwer, 1993.
[62] J. Chien, H. Wang, and C. Lee, "Improved Bayesian learning of hidden Markov models for speaker adaptation," in *Proc. ICASSP'97*, Apr. 1997, pp. 1027–1030.
[63] 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.
[64] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.
[65] Y. Sagisaka, K. Takeda, M. Abel, S. Katagiri, T. Umeda, and H. Kuwabara, "A large-scale Japanese speech database," in *Proc. ICSLP'96*, Nov. 1990, pp. 1089–1092.
[66] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, The HMM-Based Speech Synthesis System (HTS). [Online]. Available: http://www.hts.sp.nitech.ac.jp/
[67] H. Zen, T. Nose, J. Yamagishi, S. Sako, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop Speech Synth.*, Aug. 2007, pp. 294–299.



Junichi Yamagishi received the B.E. degree in computer science, and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2003, and 2006, respectively. He pioneered the use of speaker adaptation techniques in HMM-based speech synthesis in his doctoral dissertation *Average-voice-based speech synthesis*, which won the Tejima Doctoral Dissertation Award 2007.

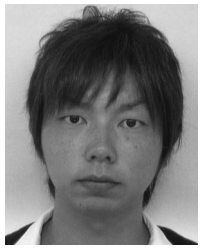
He held a research fellowship from the Japan Society for the Promotion of Science (JSPS) from 2004 to 2007. He was an Intern Researcher at ATR Spoken Language Communication Research Laboratories (ATR-SLC) from 2003 to 2006. He was a Visiting Researcher at the Center for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K. from 2006 to 2007. He is currently a Senior Research Fellow at the CSTR and continues the research on the speaker adaptation for HMM-based speech synthesis in an EC FP7 collaborative project called the EMIME project (www.emime.org). His research interests include speech synthesis, speech analysis, and speech recognition.

Dr. Yamagishi is a member of the ISCA, IEICE, and ASJ.

Takao Kobayashi (SM'04) received the B.E. degree in electrical engineering, the M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively.

In 1982, he joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, as a Research Associate. He became an Associate Professor at the same laboratory in 1989. He is currently a Professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He has served as the chair of the Speech Committee of the Institute of Electronics, Information, and Communication Engineers since 2007. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interfaces.

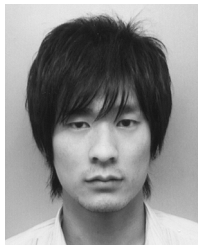
Prof. Kobayashi is a recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001 and 2008. He is a member of ISCA, IEICE, ASJ, and IPSJ.



Yuji Nakano received the B.E. degree in computer science and the M.E. degree in information processing from Tokyo Metropolitan University, Tokyo, Japan, in 2005 and 2007, respectively.

He is currently with Nissan Motor Co., Ltd., Tokyo.

Mr. Nakano received a poster award at the spring meeting held by the Acoustic Society of Japan 2006 for his study on the CSMAPLR adaptation.



Katsumi Ogata received the B.E. degree in computer science and the M.E. degree in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2005 and 2007, respectively.

He is currently with SAP Japan, Co., Ltd., Tokyo.

Juri Isogai received the B.E. degree in computer science and the M.E. degree in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 2004 and 2006, respectively.

She is currently with Nissan Motor Co., Ltd., Tokyo.