Hidden Semi-Markov Model And Its Speaker Adaptation Techniques

Junichi Yamagishi, *Member, IEEE,* Takao Kobayashi, *Member, IEEE*

I. HIDDEN SEMI-MARKOV MODEL

An N-state left-to-right HSMM λ [1], [2], [3] 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 i-th state output and duration distributions are Gaussian distributions characterized by a mean vector $\boldsymbol{\mu}_i \in \mathcal{R}^{3L}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i \in \mathcal{R}^{3L\times 3L}$, and a scalar mean m_i and variance σ_i^2, respectively; i.e.,

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

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

where $\boldsymbol{o} \in \mathcal{R}^{3L}$ is an observation vector and d is the duration in state i. The observation probability of training data $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ of length T, given the model λ, can be written as

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

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

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

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

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

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

II. CONSTRAINED MAXIMUM LIKELIHOOD LINEAR REGRESSION

Target parameters for the HSMM-based MLLR adaptation were restricted to the mean vectors of the average voice model [4]. However, we should simultaneously adapt covariance matrices to a new speaker because the covariance is also one of the important factors affecting speaker characteristics of synthetic speech. In the HMM-based CMLLR adaptation [5], mean vectors and covariance matrices of the state output pdfs are simultaneously transformed using the same linear transformation matrix (Fig. 2). Similarly, the HSMM-based CMLLR adaptation simultaneously transforms mean vectors and covariance matrices of the state output and duration pdfs using the same linear matrices as follows:

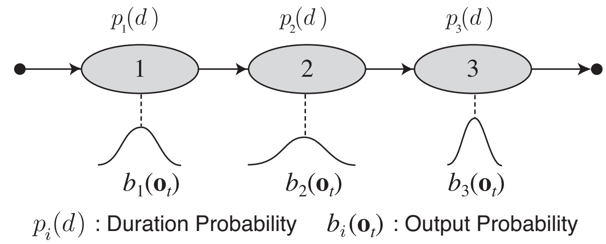

Fig. 1. Hidden Semi-Markov Model.

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

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

These transformations are equivalent to the following affine transformations of observation vector and state duration:

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

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 $\boldsymbol{\phi} = [d, 1]^\top$. $\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}] \in \mathcal{R}^{3L\times(3L+1)}$ and $\boldsymbol{X} = [\chi, \nu] \in \mathcal{R}^{1\times 2}$ are the linear transformation matrices for the state output and duration pdfs, respectively. Re-estimation formulas based on the Baum-Welch algorithm of l-th row vector $\boldsymbol{w}_l$ of $\boldsymbol{W}$ and $\boldsymbol{X}$ can be derived as follows:

$$\overline{\boldsymbol{w}}_l = (\alpha\boldsymbol{p}_l + \boldsymbol{y}_l)\boldsymbol{G}_l^{-1} \quad (15)$$

$$\overline{\boldsymbol{X}} = (\beta\boldsymbol{q} + \boldsymbol{z})\boldsymbol{K}^{-1} \quad (16)$$

where $\boldsymbol{p}_l = [0\, \boldsymbol{c}_l^\top]^\top$ and $\boldsymbol{q} = [0\, 1]^\top$. It is note that $\boldsymbol{c}_l$ is l-th cofactor row vector of $\boldsymbol{W}$. In these equations, $\boldsymbol{y}_l \in \mathcal{R}^{3L+1}$,

J. Yamagishi is with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9LW United Kingdom (see http://homepages.inf.ed.ac.uk/jyamagis/).

T. Kobayashi are with Tokyo Institute of Technology.

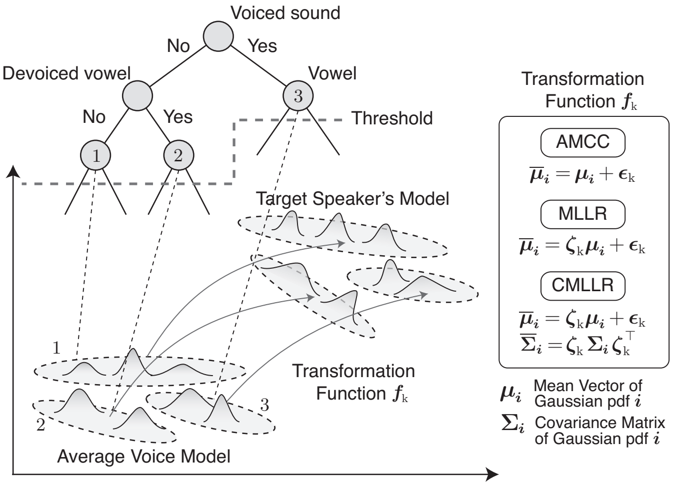

Fig. 2. Constrained Maximum Likelihood Linear Regression

$\boldsymbol{G}_l \in \mathcal{R}^{(3L+1)\times(3L+1)}$, $\boldsymbol{z} \in \mathcal{R}^2$, and $\boldsymbol{K} \in \mathcal{R}^{2\times 2}$ are given by

$$\boldsymbol{y}_l = \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\ \frac{1}{\Sigma_r(l)}\ \mu_r(l) \sum_{s=t-d+1}^{t} \boldsymbol{\xi}_s^\top \quad (17)$$

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

$$\boldsymbol{z} = \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\ \frac{1}{\sigma_r^2}\ m_r\,\boldsymbol{\phi}_s^\top \quad (19)$$

$$\boldsymbol{K} = \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\ \frac{1}{\sigma_r^2}\ \boldsymbol{\phi}_s\,\boldsymbol{\phi}_s^\top, \quad (20)$$

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

$$\alpha^2 \boldsymbol{p}_l \boldsymbol{G}_l^{-1} \boldsymbol{p}_l^\top + \alpha \boldsymbol{p}_l \boldsymbol{G}_l^{-1} \boldsymbol{y}_l^\top - \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r)\ d = 0 \quad (21)$$

$$\beta^2 \boldsymbol{q} \boldsymbol{K}^{-1} \boldsymbol{q}^\top + \beta \boldsymbol{q} \boldsymbol{K}^{-1} \boldsymbol{z}^\top - \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(r) = 0. \quad (22)$$

Since the cofactor $\boldsymbol{c}_l$ affects all row vectors of $\boldsymbol{W}$, we adopt the same updating method of $\boldsymbol{W}$ proposed in [5]. On the other hand, the estimation for $\overline{\boldsymbol{X}}$ is a closed-form. Although we explain this algortihm using global transform matrices, it is straightforward to estimate multiple transformation matrices and conduct piecewise linear regression. In order to group the distributions in the model and to tie the transformation matrices in each group, we use decision trees for context clustering in the same manner as the MLLR adaptation.

This algorithm would have effect on adaptation of prosodic information since the range of F0 and duration is one of the important factors for synthetic speech. Another advantage is that we can efficiently make the covariance matrices of the Gaussian distributions of the average voice model full matrices in the parameter generation algorithm. In [6], it is reported that full covariance modeling using semi-tied covariance [7] has effect on the parameter generation algorithm considering GV. In this system, as we can see from Eq. (7), we can use the CMLLR transform for the purpose of the full covariance modeling instead of the semi-tied covariance.

In addition to the MLLR and CMLLR adaptation, single bias removal [8], automatic model complexity control (AMCC) [9], SMAP adaptation [10], SMAPLR adaptation [11], multiple linear regression called ESAT [12] can be also used [13].

III. Feature-Space Speaker Adaptive Training

Although we utilized a model-space SAT algorithms [14] using linear transformations of mean vectors of Gaussian pdfs in our conventional systems [4], [15], a feature-space SAT algorithm [5] is used as an alternative algorithm in the AVSS 2006 system to efficiently utilize both mean vectors and covariance matrices of the Gaussian pdfs for the speaker normalization of the average voice model. We can derive the feature-space SAT in the framework of the HSMM in a similar way to [4]. The feature-space SAT of the HSMM estimates the parameters of the Gaussian pdfs as follows:

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\overline{\boldsymbol{o}}_s^{(f)}}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\ d} \quad (23)$$

$$\overline{\boldsymbol{\Sigma}}_i = \sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i) \frac{\sum_{s=t-d+1}^{t}(\overline{\boldsymbol{o}}_s^{(f)} - \overline{\boldsymbol{\mu}}_i)(\overline{\boldsymbol{o}}_s^{(f)} - \overline{\boldsymbol{\mu}}_i)^\top}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\ d} \quad (24)$$

$$\overline{m}_i = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\ \overline{d}^{(f)}}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)} \quad (25)$$

$$\overline{\sigma}_i^2 = \frac{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)\ (\overline{d}^{(f)} - \overline{m}_i)^2}{\sum_{f=1}^{F}\sum_{t=1}^{T_f}\sum_{d=1}^{t}\gamma_t^d(i)} \quad (26)$$

where F is number of the training speakers and T_f is total number of frames of a speaker f. Note that $\overline{\boldsymbol{o}}_s = \overline{\boldsymbol{\zeta}}\boldsymbol{o}_s + \overline{\boldsymbol{\epsilon}}$ and $\overline{d} = \overline{\chi} d + \overline{\nu}$ are linearly transformed observation vector and duration in the framework of the HSMM-based CMLLR adaptation. This technique can be viewed as a generalized version of several normalization techniques such as cepstral mean normalization (CMN) [16], cepstral variance normalization (CVN) [17], [18], vocal tract length normalization (VTLN) [19], [20], and bias removal of F_0 and duration. Since this HSMM-based feature-space SAT algorithm requires a lot of computations, we basically train the acoustic models using the HMM-based feature-space SAT algorithm and apply the HSMM-based SAT algorithm in the final embedded training procedures.

Another advantage of this feature-space SAT is feasibility. As reported in [5], in the the model-space SAT algorithms, it is necessary to store a full matrix for each Gaussian pdf, or store statistics for each Gaussian component for every speaker. In our *speaker-independent* HMM-based speech synthesis system, the number of the Gaussian pdfs reaches $\mathcal{O}(10^7)$ or more, and it partly makes the parameter estimation impractical.

In particular, the embedded training procedures in which we could use the model-space SAT were restricted to the training procedures in which the parameters of the Gaussian pdfs were tied among several pdfs. On the other hand, we can apply the feature-space SAT algorithm to all the embedded training procedures and conduct further normalization in the training of the average voice model.

IV. Constrained Structural Maximum A Posteriori Linear Regression

The CMLLR adaptation algorithm utilizes the maximum likelihood criterion for the estimation of the transformation matrices. In the training stage of the average voice model using the SAT algorithm, the criterion would work well since large amount of training data for the average voice model is available. However, in the adaptation stage, the amount of adaptation data is very limited. Hence, we need to use more robust criteria such as maximum a posteriori criterion. In the MAP estimation, we estimate the transformation matrices as follows:

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \boldsymbol{W}) P_b(\boldsymbol{W}) \tag{27}$$

$$\widehat{\boldsymbol{X}} = \underset{\boldsymbol{X}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \boldsymbol{X}) P_p(\boldsymbol{X}) \tag{28}$$

where $P_b(\boldsymbol{W})$ and $P_p(\boldsymbol{X})$ is priori distributions for the transformation matrices $\boldsymbol{W}$ and $\boldsymbol{X}$, respectively. For the prior distributions, the following matrix variate normal distributions, matrix versions of the multivariate normal distribution [21] are convenient:

$$P_b(\boldsymbol{W}) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} \exp\Big\{-\frac{1}{2} tr(\boldsymbol{W}-\boldsymbol{H})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{W}-\boldsymbol{H})\boldsymbol{\Psi}^{-1}\Big\} \tag{29}$$

$$P_p(\boldsymbol{X}) \propto |\tau_p|^{-1} |\boldsymbol{\psi}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2} tr(\boldsymbol{X}-\boldsymbol{\eta})^\top \tau_p^{-1}(\boldsymbol{X}-\boldsymbol{\eta})\boldsymbol{\psi}^{-1}\Big\} \tag{30}$$

where $\boldsymbol{\Omega} \in \mathcal{R}^{3L\times 3L}$, $\boldsymbol{\Psi} \in \mathcal{R}^{(3L+1)\times(3L+1)}$, $\boldsymbol{H} \in \mathcal{R}^{3L\times(3L+1)}$, $\tau_p > 0$, $\boldsymbol{\psi} \in \mathcal{R}^{2\times 2}$, and $\boldsymbol{\eta} \in \mathcal{R}^{1\times 2}$ are the hyperparameters for the prior distributions.

In the SMAP criterion [10], tree structures of the distributions effectively cope with the control of the hyperparameters. Specifically, we first estimate global transformation parameters at a rood node of the tree structure using all the adaptation data, and then propagate it to its child nodes as their hyperparameters $\boldsymbol{H}$ and $\boldsymbol{\eta}$. In the child nodes, the transformation matrices are estimated again using their adaptation data, based on the MAP criterion with the propagated hyperparameters. Then, the recursive MAP-based estimation of the transformation matrices from a root node to lower nodes is conducted (Fig. 3). Shiohan *et al.* applied the SMAP criterion to the MLLR and developed SMAPLR adaptation [11].

In this paper, we apply the SMAP criterion to the CMLLR adaptation, and estimate the transformation matrices for simultaneously transforming mean vectors and covariance matrices of state output and duration distributions using the recursive MAP criterion. This algorithm is called "constrained structural maximum a posteriori linear regression," or for short,

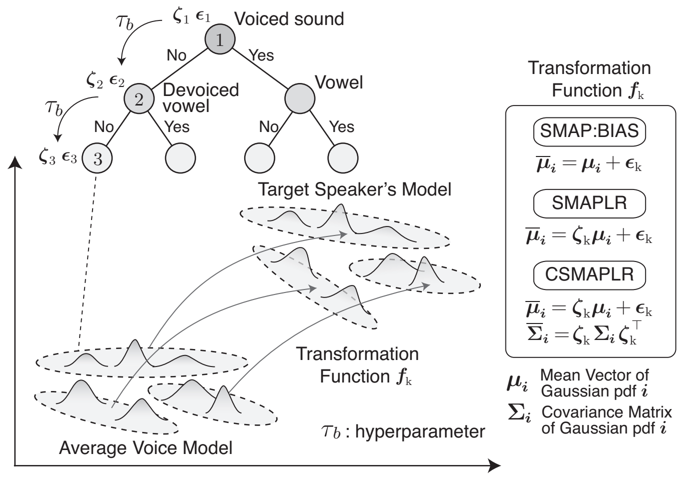

Fig. 3. Constrained Structural Maximum A Posteriori Linear Regression

"CSMAPLR". In the CSMAPLR adaptation, we fix $\boldsymbol{\Psi}$ and $\boldsymbol{\psi}$ to the identity matrices, and set $\boldsymbol{\Omega}$ to a scaled identity matrix $\boldsymbol{\Omega} = \tau_b \boldsymbol{I}_{3L}$ so that the scaling is controlled by a positive scalar coefficient τ_b in the same manner as SMAPLR adaptation [11]. Here $\boldsymbol{I}_{3L}$ is the $3L \times 3L$ identity matrix. We use the same notation method for different dimensional identity matrices. Re-estimation formulas based on the Baum-Welch algorithm of the transformation matrices can be derived as follows:

$$\widehat{\boldsymbol{w}}_l = (\alpha \boldsymbol{p}_l + \boldsymbol{y}'_l)\boldsymbol{G}'^{-1}_l \tag{31}$$

$$\widehat{\boldsymbol{X}} = (\beta \boldsymbol{q} + \boldsymbol{z}')\boldsymbol{K}'^{-1}. \tag{32}$$

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

$$\boldsymbol{y}'_l = \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\Sigma_r(l)}\, \mu_r(l) \sum_{s=t-d+1}^{t} \boldsymbol{\xi}_s^\top + \tau_b\, \boldsymbol{h}_l \tag{33}$$

$$= \boldsymbol{y}_l + \tau_b\, \boldsymbol{h}_l \tag{34}$$

$$\boldsymbol{G}'_l = \sum_{r=1}^{R_b}\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\Sigma_r(l)} \sum_{s=t-d+1}^{t} \boldsymbol{\xi}_s \boldsymbol{\xi}_s^\top + \tau_b\, \boldsymbol{I}_{3L+1} \tag{35}$$

$$= \boldsymbol{G}_l + \tau_b\, \boldsymbol{I}_{3L+1} \tag{36}$$

$$\boldsymbol{z}' = \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\sigma_r^2}\, m_r\, \boldsymbol{\phi}_r^\top + \tau_p\, \boldsymbol{\eta} \tag{37}$$

$$= \boldsymbol{z} + \tau_p\, \boldsymbol{\eta} \tag{38}$$

$$\boldsymbol{K}' = \sum_{r=1}^{R_p}\sum_{t=1}^{T}\sum_{d=1}^{t} \gamma_t^d(r)\, \frac{1}{\sigma_r^2}\, \boldsymbol{\phi}_r\, \boldsymbol{\phi}_r^\top + \tau_p\, \boldsymbol{I}_2 \tag{39}$$

$$= \boldsymbol{K} + \tau_p\, \boldsymbol{I}_2 \tag{40}$$

where $\boldsymbol{h}_l$ is the l-th row vector of $\boldsymbol{H}$. The quadratic equations for α and β are the same as Eqs. (21) and (22).

The CSMAPLR adaptation algorithm can utilize the tree structure more effectively than the CMLLR adaptation since the tree structure represents connection and similarity between the distributions, and the propagated prior information automatically reflects the connection and similarity. Additionally, our tree structures used in these experiments represent linguistic information as shown in Figs. 3.Hence, the propagated prior information would reflect the connection and similarity of the distributions in keeping with the linguistic information.

REFERENCES

[1] J. Ferguson, "Variable duration models for speech," in *Symp. on the Application of Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.

[2] 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.

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

[4] 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.

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

[6] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," in *Proc. Blizzard Challenge 2006*, Sep. 2006.

[7] M. Gales, "Semi-tied covariance matrices for hidden Markvo models," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 272–281, Mar. 1999.

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

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

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

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

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

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

[14] 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.

[15] 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 2007*, Apr. 2007, pp. 1233–1236.

[16] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.

[17] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[18] C. Chen, K. Filali, and J. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *Proc. ICSLP 2002*, Sep. 2002, pp. 241–244.

[19] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP 1996*, May 1999, pp. 346–348.

[20] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP 1996*, May 1999, pp. 353–356.

[21] A. Gupta and T. Varga, *Elliptically Contoured Models in Statistics*. Kluwer Academic Publishers, 1993.

PLACE PHOTO HERE

Junichi Yamagishi received the B.E. degree in computer science, M.E. and Dr.Eng. degrees in information processing from 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) during 2004 - 2007. He was an intern researcher at ATR spoken language communication Research Laboratories (ATR-SLC) during 2003 - 2006. He was a visiting researcher at the Centre for Speech Technology Research (CSTR), University of Edinburgh, U.K. during 2006 - 2007. He is currently a senior research fellow at the CSTR, University of Edinburgh. His research interests include speech synthesis, speech analysis, and speech recognition. He is a member of IEEE, ISCA and ASJ.

PLACE PHOTO HERE

Takao Kobayashi 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 is a co-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. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interface. He is a member of IEEE, ISCA, IPSJ and ASJ.