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

Source: TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, JANUARY 2009

Author : Junichi Yamagishi, Takao Kobayashi

Professor : 陳嘉平

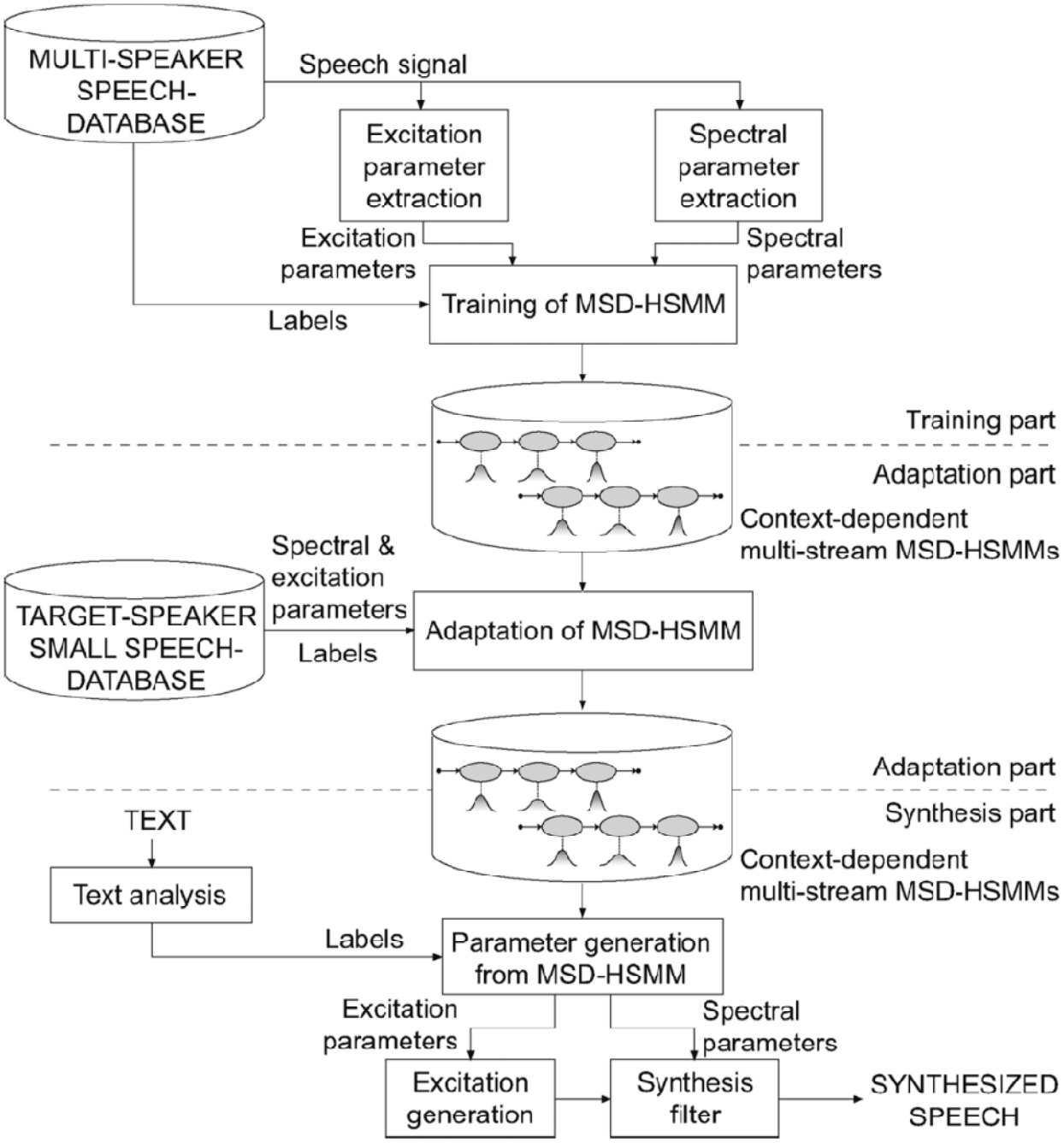
Reporter : 楊治鏞

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.

Introduction

- The maximum-likelihood 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.*


MULTI-SPEAKER SPEECH-DATABASE
Speech signal
Excitation parameter extraction
Spectral parameter extraction
Excitation parameters
Spectral parameters
Training of MSD-HSMM
Labels
Training part
Adaptation part
Context-dependent multi-stream MSD-HSMMs
Spectral & excitation parameters
TARGET-SPEAKER SMALL SPEECH-DATABASE
Adaptation of MSD-HSMM
Labels
Adaptation part
Synthesis part
Context-dependent multi-stream MSD-HSMMs
TEXT
Text analysis
Labels
Parameter generation from MSD-HSMM
Excitation parameters
Spectral parameters
Excitation generation
Synthesis filter
SYNTHESIZED SPEECH

Speaker Adaptation Based on HSMM

- We assume that the i -th state output and duration distributions are Gaussian distributions characterized by a mean vector $\boldsymbol{\mu}_i \in \mathcal{R}^L$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i \in \mathcal{R}^{L \times L}$ and a scalar mean m_i and variance σ_i^2 .

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$
$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2)$$

- where $\boldsymbol{o} \in \mathcal{R}^L$ is an observation vector and d is the duration in state i.

CSMAPLR

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

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

Constrained Structural Maximum A Posteriori Linear Regression

- In the MAP estimation, we estimate the transforms as follows:

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

- where $P(\Lambda)$ is a prior distribution for the transforms $\boldsymbol{W}$ and $\boldsymbol{X}$.

Constrained Structural Maximum A Posteriori Linear Regression

- For the prior distribution, the following combined matrix variate normal distributions are convenient:

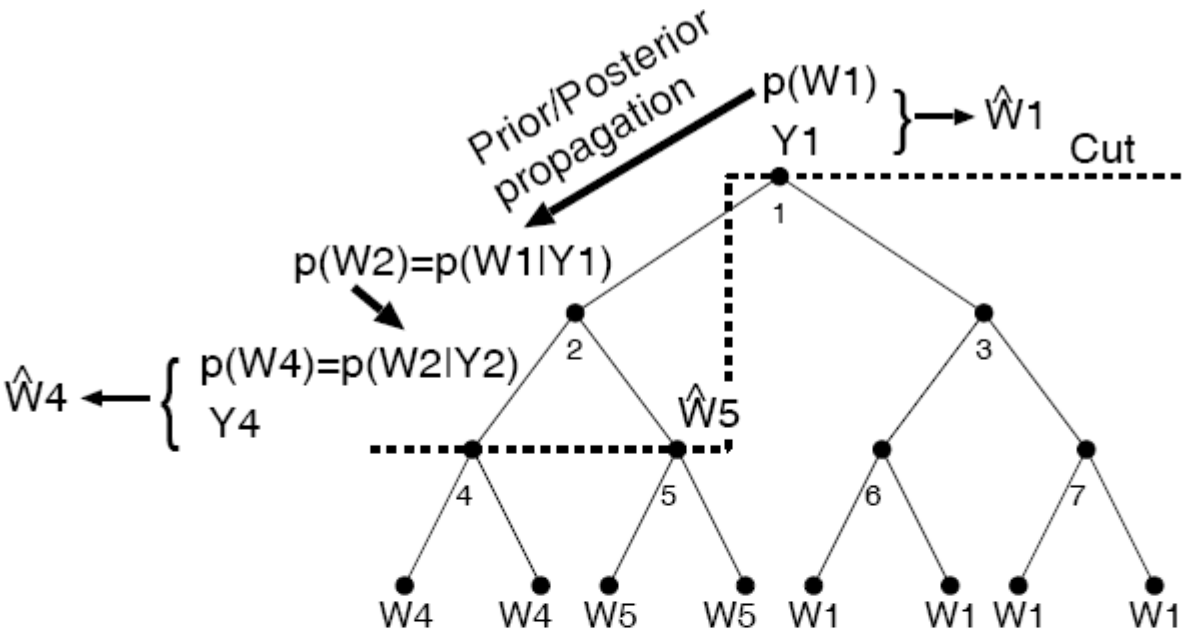
$$P(\Lambda) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} |\tau_p|^{-1} |\boldsymbol{\psi}|^{-\frac{1}{2}}$$
$$\times \exp\left\{ -\frac{1}{2} tr(\boldsymbol{W} - \boldsymbol{H})^\top \boldsymbol{\Omega}^{-1} (\boldsymbol{W} - \boldsymbol{H}) \boldsymbol{\Psi}^{-1}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

- 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.

SMAP

- 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 H and η .
- 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.



Estimation in node i along the cut:

Wi = argmax p(Yi | Wi,Λ) p(Wi)

with p(Wi) = p(Wj | Yj) where j is the parent node of i

Figure 3: Tree-based SMAPLR algorithm. The adaptation data associated to a node i is denoted $\mathbf{Y}_i$. The corresponding prior density is denoted $p(\mathbf{W}_i)$. For each node i of parent j, the prior density $p(\mathbf{W}_i)$ is de□ned as $p(\mathbf{W}_j|\mathbf{Y}_j)$.

CSMAPLR

- In CSMAPLR adaptation, we fix $\boldsymbol{\Psi}$ and ψ 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 .

CSMAPLR

- Re-estimation formulas:

$$\widehat{\boldsymbol{w}}_l = \left(\alpha \boldsymbol{p}_l + \boldsymbol{y}'_l\right) \boldsymbol{G}'^{-1}_l$$
$$\widehat{\boldsymbol{X}} = \left(\beta \boldsymbol{q} + \boldsymbol{z}'\right) \boldsymbol{K}'^{-1}$$

- where $\widehat{\boldsymbol{w}}_l$ is the l-th row vector of $\boldsymbol{W}$, $\boldsymbol{p}_l = [0, \boldsymbol{c}_l]$, $\boldsymbol{q} = [0, 1]$, and $\boldsymbol{c}_l$ is the l-th cofactor row vector of $\boldsymbol{W}$.

CSMAPLR

- Then $\boldsymbol{y}'_l$, $\boldsymbol{G}'_l$, $\boldsymbol{z}'$, and $\boldsymbol{K}'$ are given by

$$\begin{aligned}
\boldsymbol{y}'_l &= \boldsymbol{y}_l + \tau_b \boldsymbol{h}_l \\
\boldsymbol{G}'_l &= \boldsymbol{G}_l + \tau_b \boldsymbol{I}_{L+1} \\
\boldsymbol{z}' &= \boldsymbol{z} + \tau_p \boldsymbol{\eta} \\
\boldsymbol{K}' &= \boldsymbol{K} + \tau_p \boldsymbol{I}_2
\end{aligned}$$

- where $\boldsymbol{h}_l$ is the lth row vector of $\boldsymbol{H}$.

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

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

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

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

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

- where $\Sigma_r(l)$ is the l-th diagonal element of the diagonal covariance matrix Σ_r, and $\mu_r(l)$ is the l-th element of the observation vector $\boldsymbol{\mu}_r$.

CSMAPLR

- Then α and β are scalar values which satisfy the following quadratic equations:

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

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

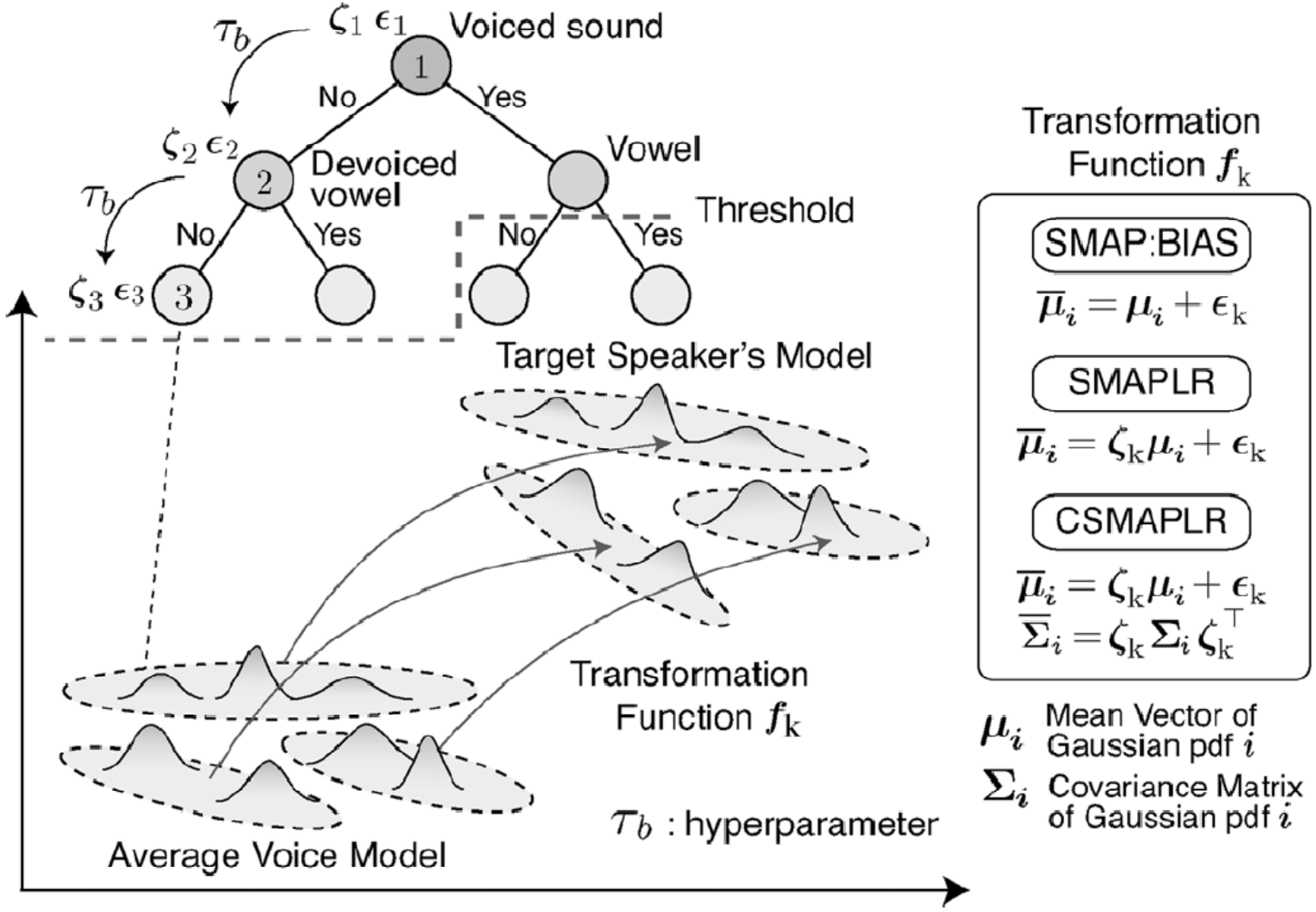


Fig. 5. Constrained structural maximum *a posteriori* linear regression (CSMAPLR) and its related algorithms.

Experimental Conditions

- We used the ATR Japanese speech database (Set B), which contains a set of 503 phonetically balanced sentences uttered by 6 male speakers and 4 female speakers.

- 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.

Evaluation of Transform Functions and Estimation Criteria

- 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 $\log F_0$.

- 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.

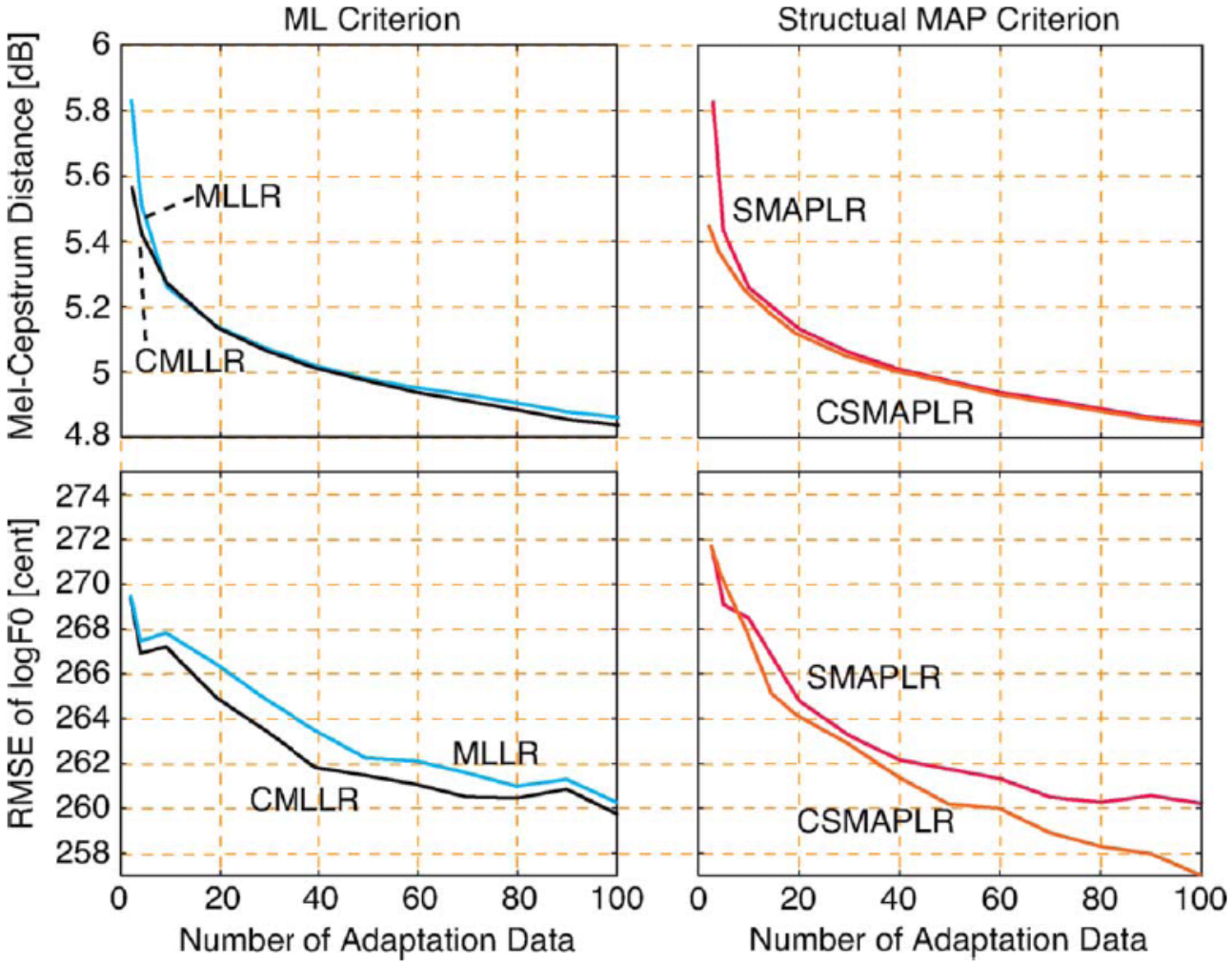


Fig. 8. Objective evaluation of transform functions in linear regression algorithms. Upper: average Mel-cepstral distance [dB], lower: RMSE of $\log F_0$ [cent].

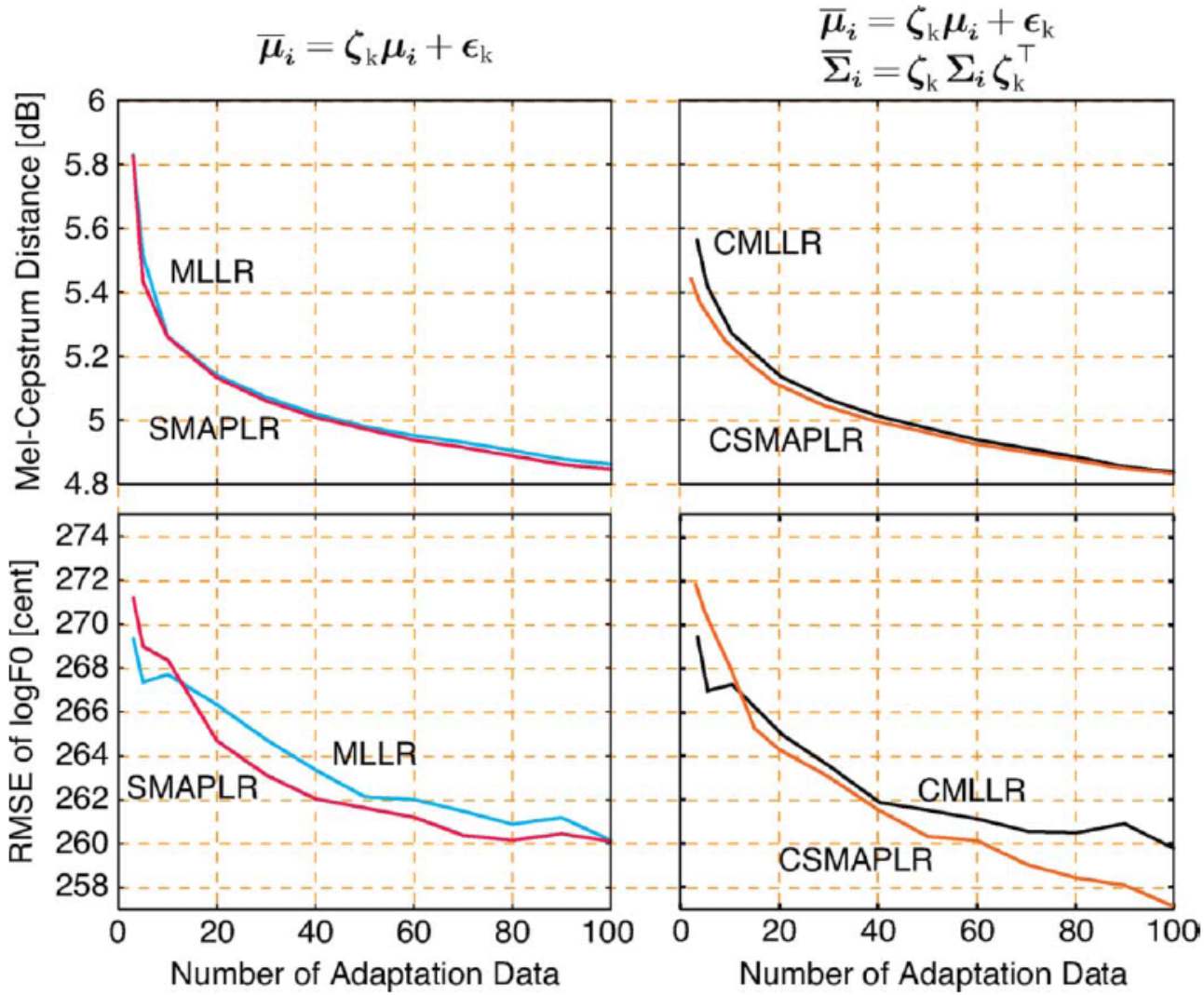


Fig. 9. Objective evaluation of estimation criteria in linear regression algorithms. Upper: average Mel-cepstral distance [dB], lower: RMSE of $\log F_0$ [cent].

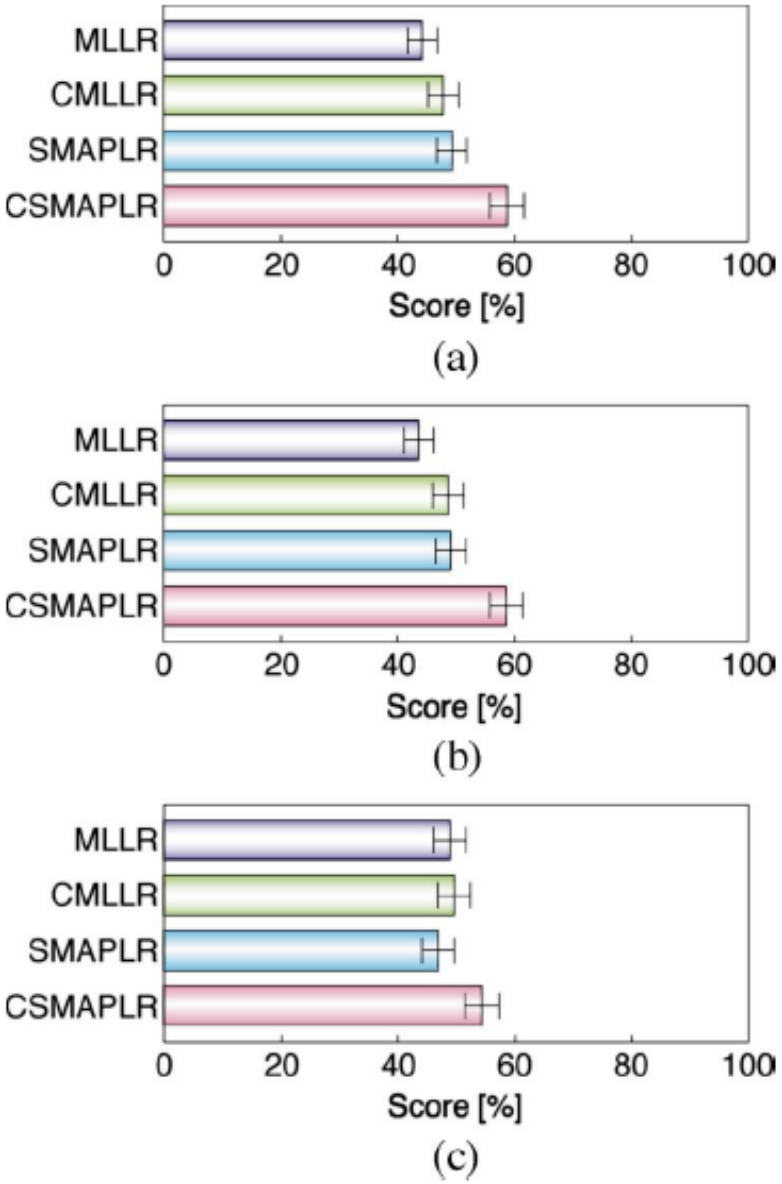
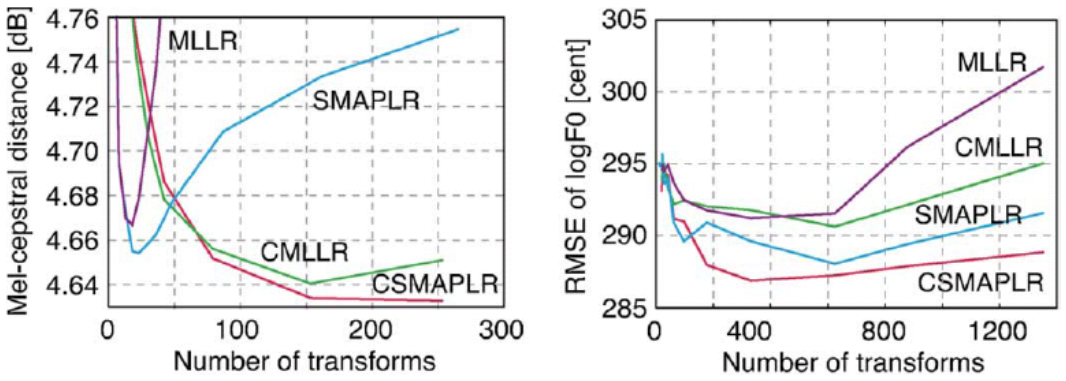


Fig. 10. Subjective evaluation of the similarity of synthetic speech generated from models adapted using several linear regression algorithms. (There were 50 adaptation sentences.) (a) Target speaker MHT. (b) Target speaker MTK. (c) Target speaker MMI.

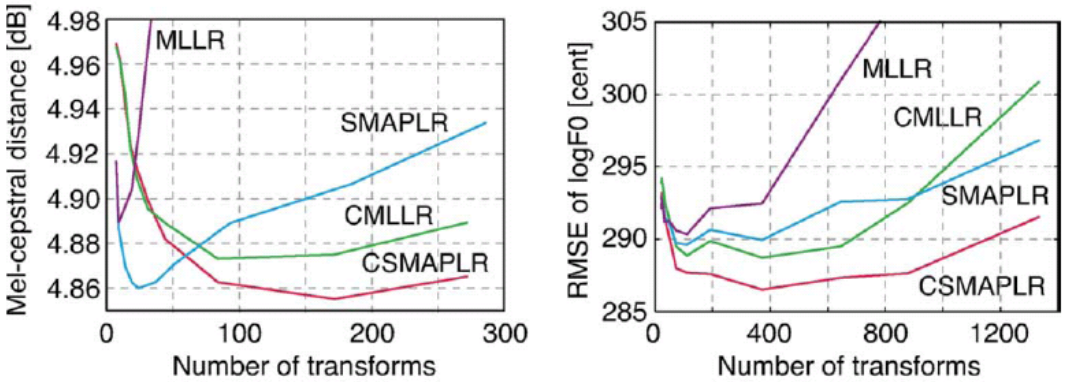
Evaluation of Sensitivity

- We gradually increased the number of the transforms for each algorithm and calculated the objective measures.

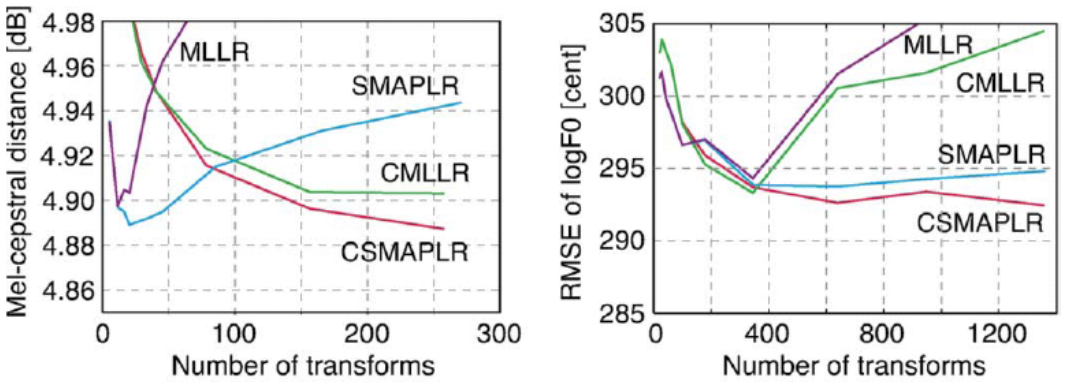
- When the number of transforms increases more than necessary, however, the amount of adaptation data used for estimating a single transform relatively decreases and over-fitting to the adaptation data occurs.


Mel-cepstral distance [dB]
4.76
4.74
4.72
4.70
4.68
4.66
4.64
MLLR
SMAPLR
CMLLR
CSMAPLR
0
100
200
300
Number of transforms
RMSE of logF0 [cent]
305
300
295
290
285
MLLR
CMLLR
SMAPLR
CSMAPLR
0
400
800
1200
Number of transforms

(a)


Mel-cepstral distance [dB]
4.98
4.96
4.94
4.92
4.90
4.88
4.86
MLLR
SMAPLR
CMLLR
CSMAPLR
0
100
200
300
Number of transforms
RMSE of logF0 [cent]
305
300
295
290
285
MLLR
CMLLR
SMAPLR
CSMAPLR
0
400
800
1200
Number of transforms

(b)


Mel-cepstral distance [dB]
4.98
4.96
4.94
4.92
4.90
4.88
4.86
MLLR
SMAPLR
CMLLR
CSMAPLR
0
100
200
300
Number of transforms
RMSE of logF0 [cent]
305
300
295
290
285
MLLR
CMLLR
SMAPLR
CSMAPLR
0
400
800
1200
Number of transforms

(c)

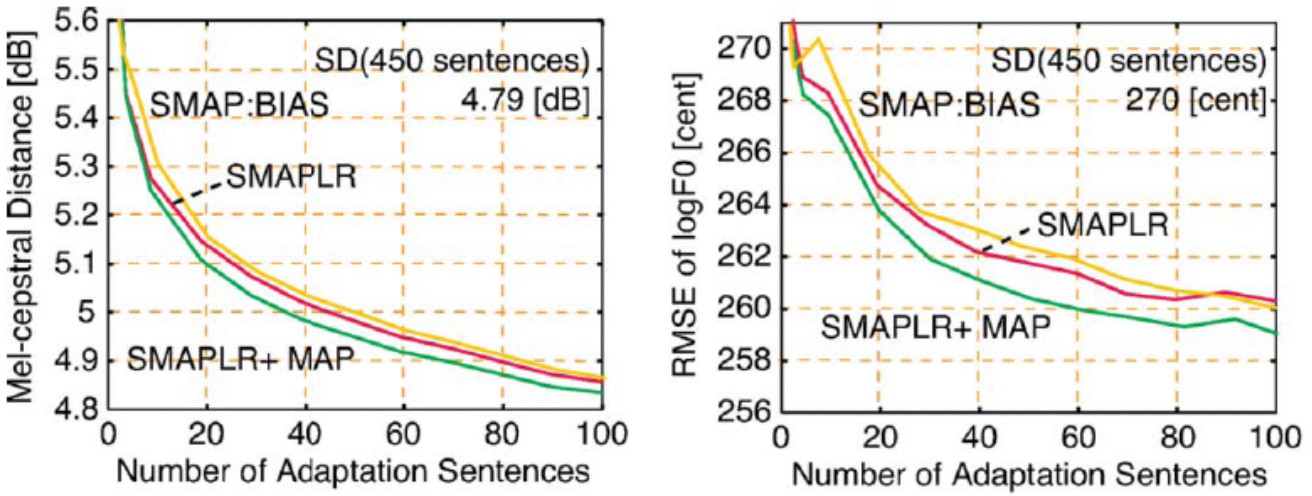
Combined Algorithm With Linear Regression and MAP Adaptation

- The MAP adaptation of mean vectors of the Gaussian pdfs transformed by the CSMAPLR algorithm can be simply estimated as follows:

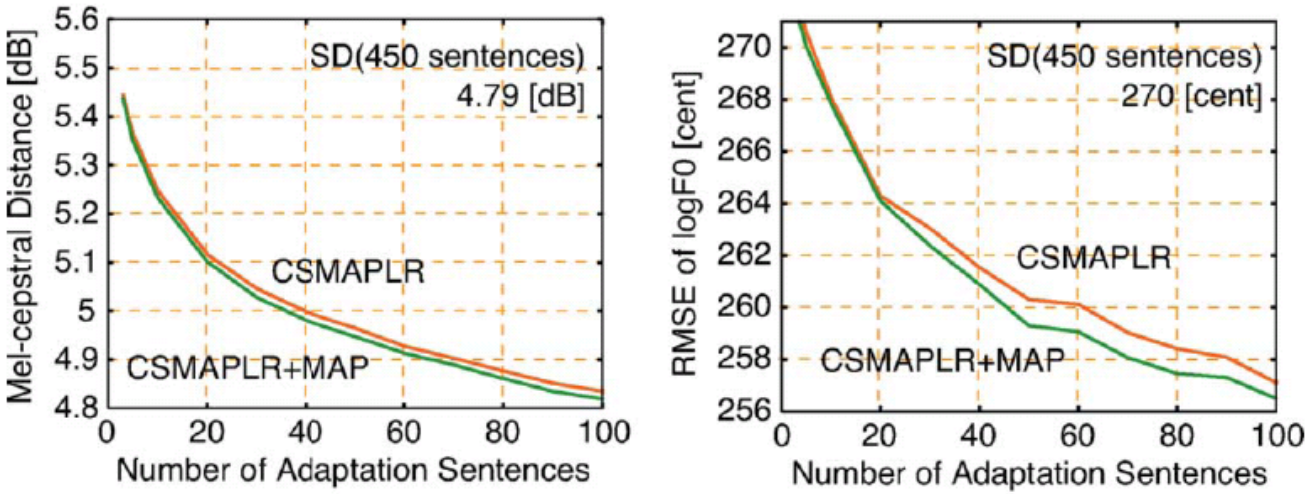
$$\widehat{\boldsymbol{\mu}}_i = \frac{\upsilon_b \boldsymbol{\mu}_i + \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(i) \sum_{s=t-d+1}^{t} \widehat{\boldsymbol{o}}_s}{\upsilon_b + \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(i) d}$$

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

- Where $\widehat{\boldsymbol{o}}_s = \widehat{\boldsymbol{\zeta}} \boldsymbol{o}_s + \widehat{\boldsymbol{\epsilon}}$ and $\widehat{d} = \widehat{\chi} d + \widehat{\nu}$ are linearly transformed observation vector and duration using the HSMM-based CSMAPLR adaptation.



(a)



(b)

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.

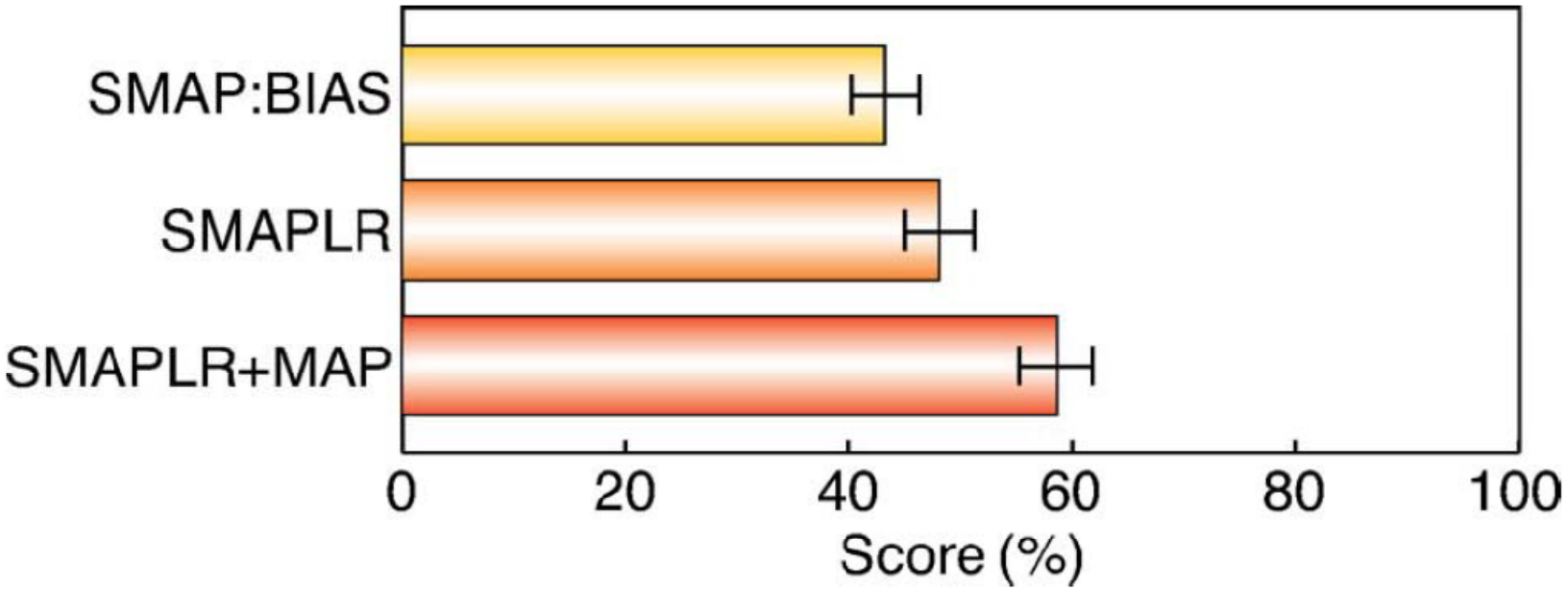


Fig. 13. Subjective evaluation of SMAP-based linear regression algorithms and an algorithm combining MAP adaptation with SMAP-based linear regression. (There were 50 adaptation sentences.)

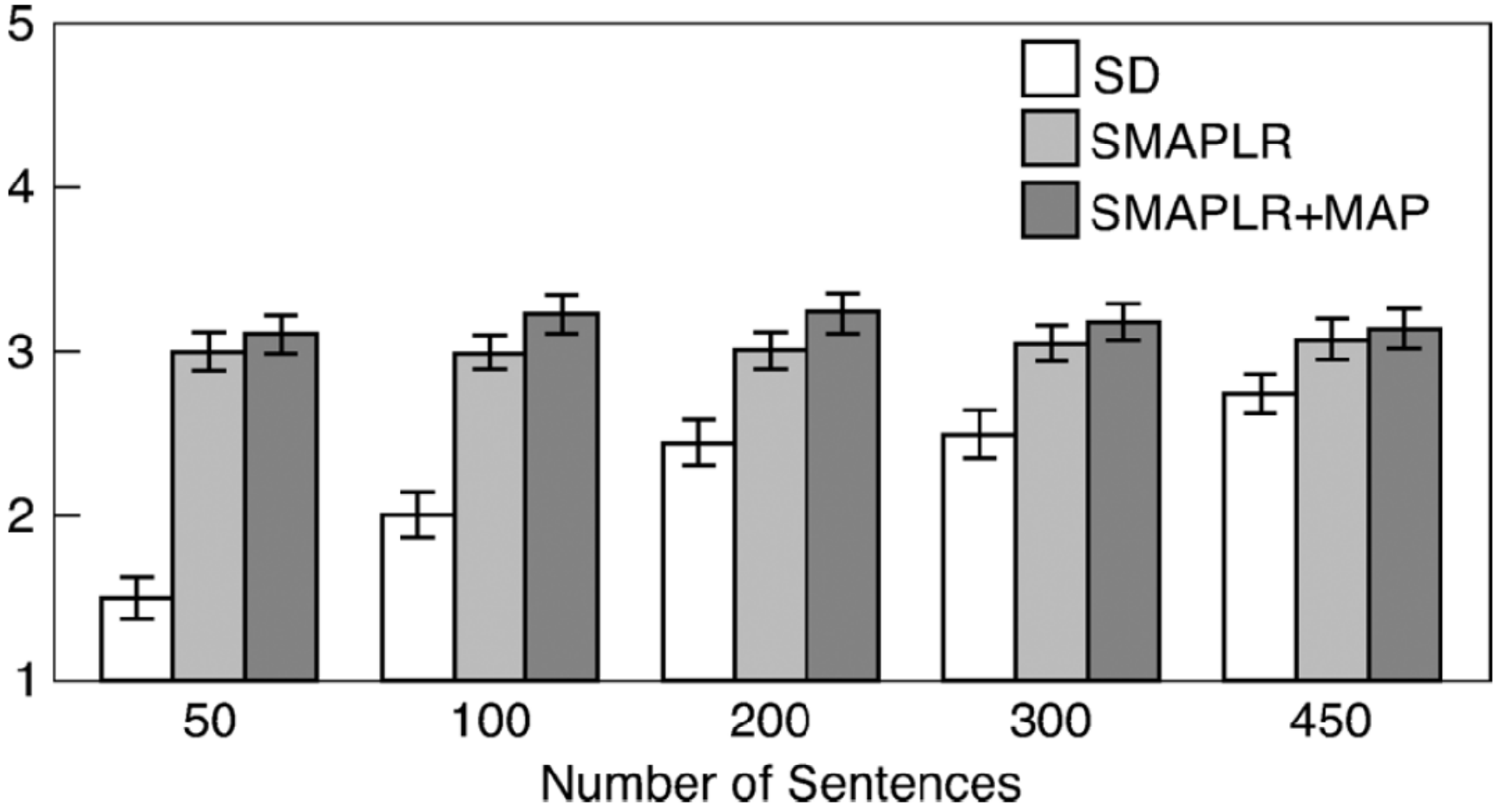


Fig. 14. Subjective evaluation of models adapted using SMAPLR or using SMAPLR combined with MAP adaptation. Similarity was rated on a 5-point scale: 5 for very similar, 4 for similar, 3 for slightly similar, 2 for dissimilar, and 1 for very dissimilar.