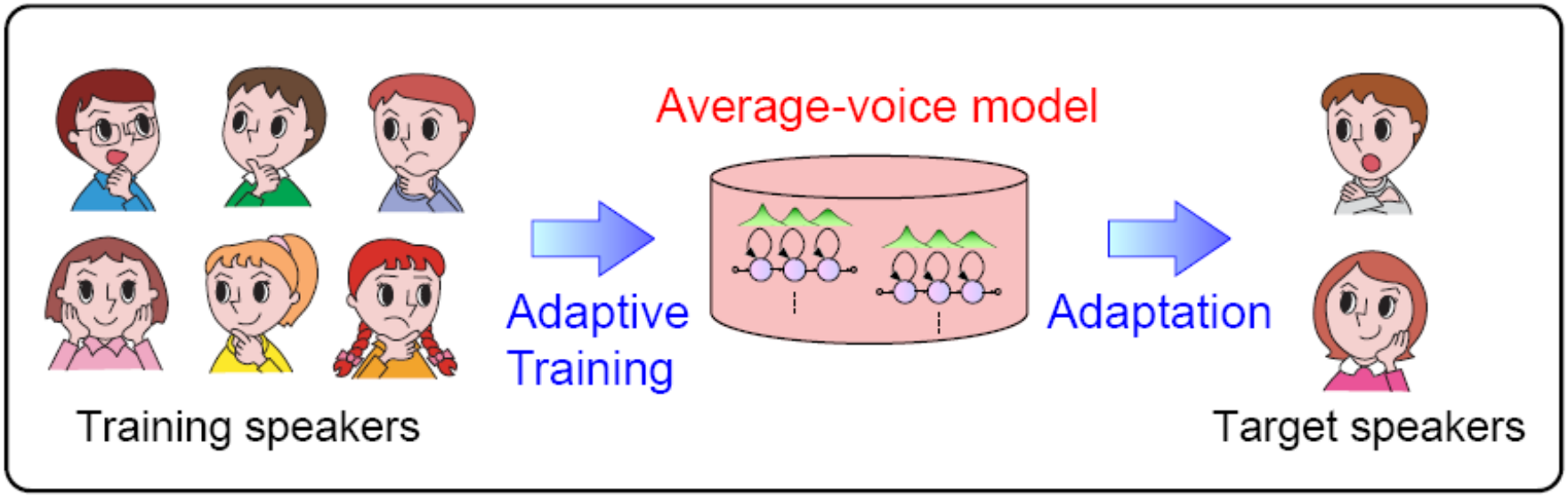
Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training

Source: IEICE TRANS. INF. & SYST.
Author : Junichi Yamagishi, Takao Kobayashi
Professor : 陳嘉平
Reporter : 楊治鏞

Speaker Adaptation

- Originally developed in ASR, but works very well in TTS
- Average voice-based speech synthesis (AVSS) [Yamagishi;'06]



- Require small data of target speaker/speaking style

⇒ **Smal cost to create new voices**

Introduction

- In the previous work, the HMM-based speaker adaptation and speaker adaptive training were conducted for transforming and normalizing only state output probability distributions corresponding to spectrum and F0 parameters of the speech data.

- The HSMM is an HMM with explicit state-duration probability distributions and enables us to conduct simultaneous adaptation of the output distributions and state duration distributions.

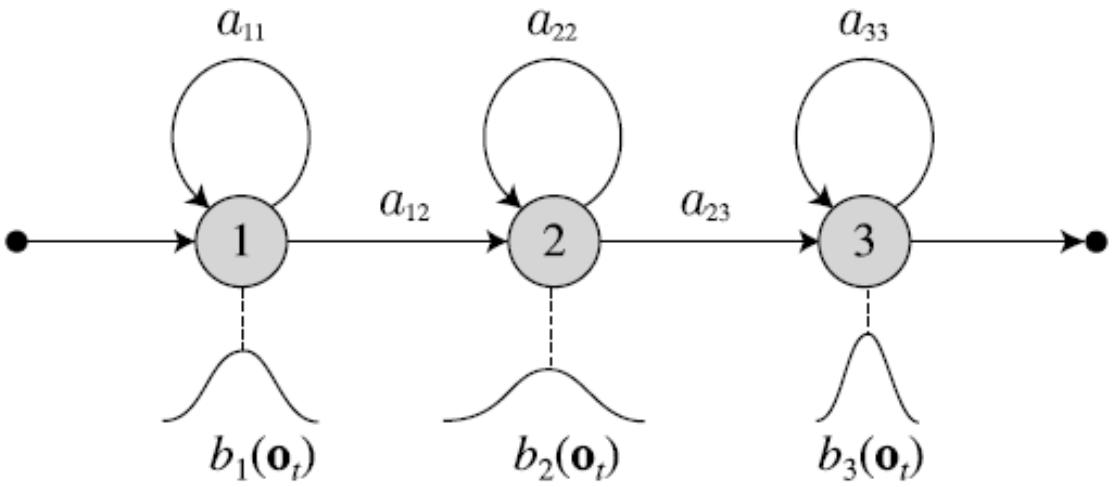


Fig. 1 Hidden Markov model.

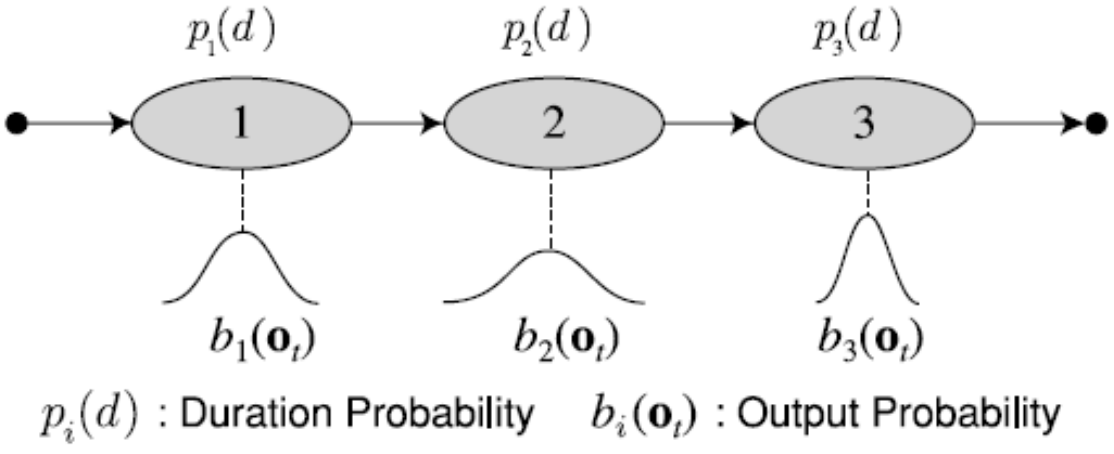
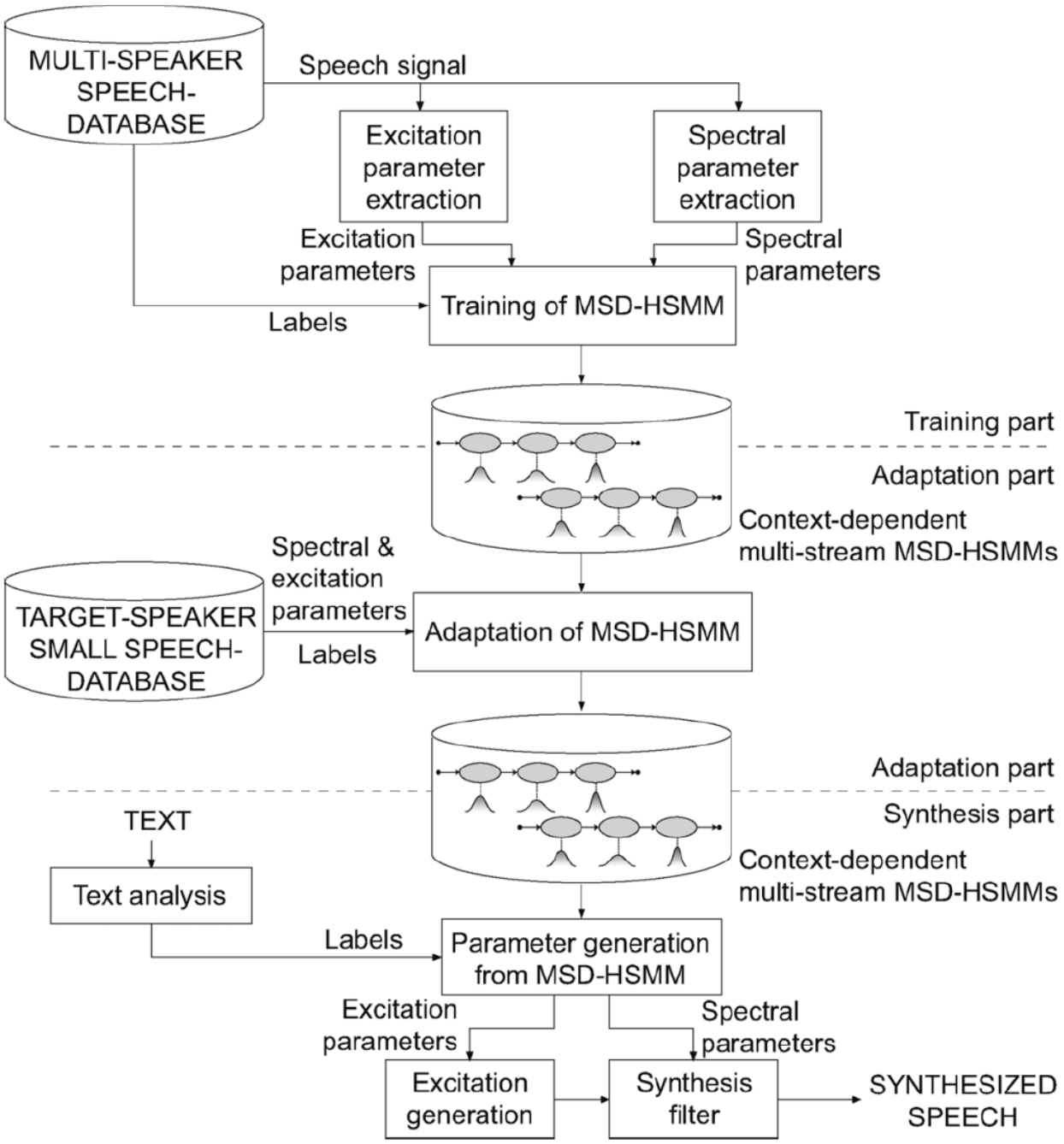


Fig. 2 Hidden semi-Markov model.


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

Hidden Semi-Markov Model

- 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 $\boldsymbol{o} \in \mathcal{R}^{3L}$ is an observation vector and d is the duration in state i .

Hidden Semi-Markov Model

- The observation probability of training data $O = (o_1, \cdots, o_T)$ of length T , given the model λ, can be written as

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

- Where $\forall\ t \in [1, T]$.

Hidden Semi-Markov Model

- 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

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

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

- In the HSMM-based MLLR adaptation, the mean vectors of the state output and duration distributions for the target speaker are obtained by linearly transforming the mean vectors of the state output and duration distributions of the average voice model (Fig. 3) as follows:

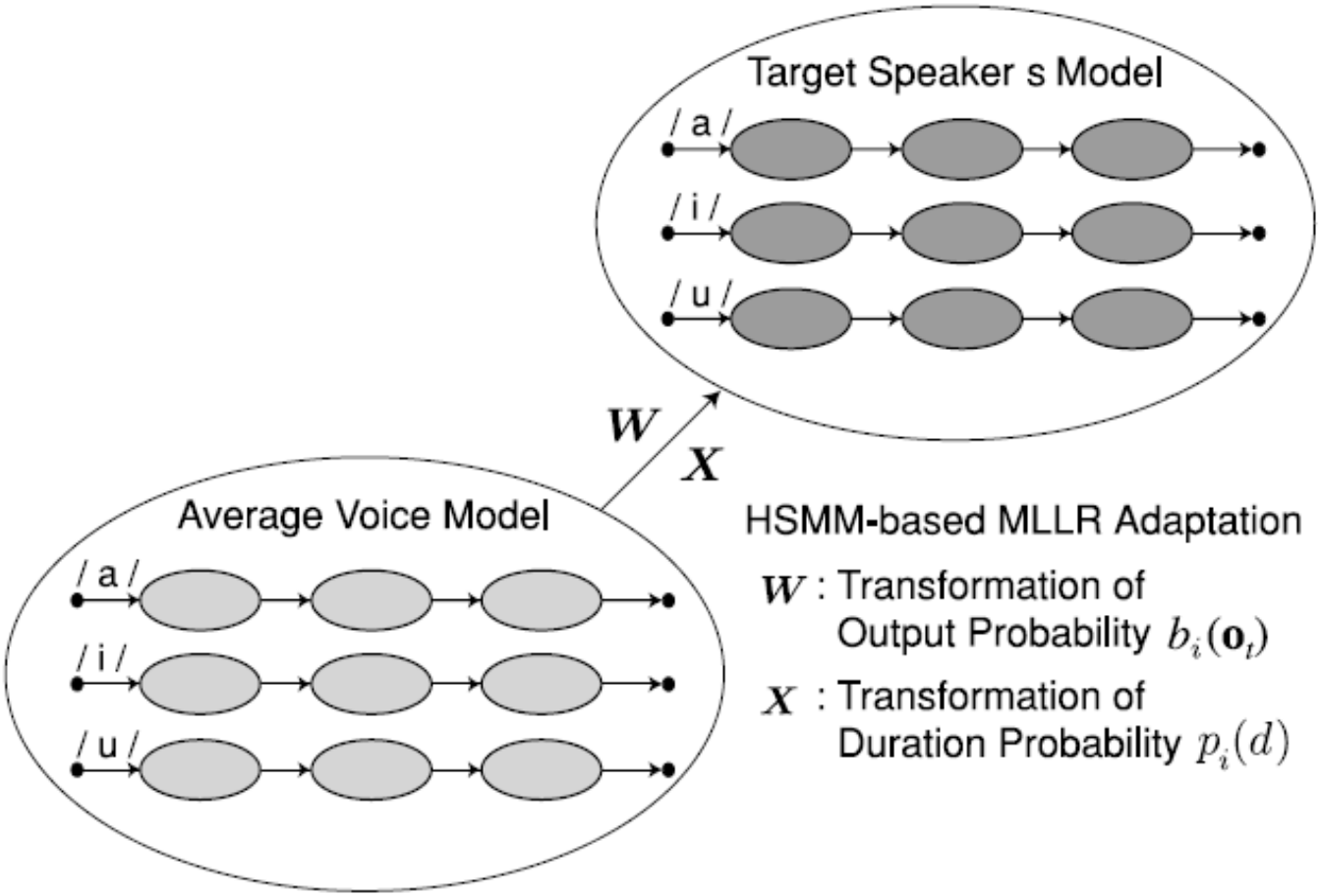


Fig. 3 HSMM-based MLLR adaptation.

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

- $$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\zeta}\boldsymbol{\mu}_i + \boldsymbol{\epsilon}, \boldsymbol{\Sigma}_i)$$
$$= \mathcal{N}(\boldsymbol{o}; \boldsymbol{W}\boldsymbol{\xi}_i, \boldsymbol{\Sigma}_i)$$
$$p_i(d) = \mathcal{N}(d; \chi m_i + \nu, \sigma_i^2)$$
$$= \mathcal{N}(d; \boldsymbol{X}\boldsymbol{\phi}_i, \sigma_i^2)$$

- $\boldsymbol{W} = [\boldsymbol{\zeta}, \boldsymbol{\epsilon}] \in \mathcal{R}^{L\times(L+1)}$ and $\boldsymbol{X} = [\chi, \nu] \in \mathcal{R}^{1\times 2}$ are the transformation matrices which transform extended mean vectors $\boldsymbol{\xi}_i = [\boldsymbol{\mu}_i^\top, 1]^\top \in \mathcal{R}^{L+1}$ and $\boldsymbol{\phi}_i = [m_i, 1]^\top \in \mathcal{R}^2$, respectively.

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

- The HSMM-based MLLR adaptation estimates a set of transformation matrices $\Lambda = (W, X)$ so that the likelihood of the adaptation data O of length T is maximized.

- The problem of the HSMM-based MLLR adaptation based on the ML criterion can be expressed as follows:

$$\tilde{\Lambda} = \left(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}\right) = \underset{\Lambda}{\operatorname{argmax}}\, P(\boldsymbol{O}|\lambda, \Lambda)$$

- where λ is the parameter set of HSMM.

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

- Re-estimation formulas:

$$\overline{\boldsymbol{w}}_l = \boldsymbol{y}_l \boldsymbol{G}_l^{-1}$$
$$\overline{\boldsymbol{X}} = \boldsymbol{z}\boldsymbol{K}^{-1}$$

- where $\boldsymbol{w}_l \in \mathcal{R}^{L+1}$ is the l-th row vector of W.
- In these equations, $\boldsymbol{y}_l \in \mathcal{R}^{L+1}$, $\boldsymbol{G}_l \in \mathcal{R}^{(L+1)\times(L+1)}$, $\boldsymbol{z} \in \mathcal{R}^2$, and $\boldsymbol{K} \in \mathcal{R}^{2\times 2}$ are given by

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

$$\boldsymbol{y}_l = \sum_{r=1}^{R_b} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r) \, \frac{1}{\Sigma_r(l)} \sum_{s=t-d+1}^{t} o_s(l) \, \boldsymbol{\xi}_r^{\top} \qquad \boldsymbol{z} = \sum_{r=1}^{R_p} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r) \, \frac{1}{\sigma_r^2} \, d \, \boldsymbol{\phi}_r^{\top}$$

$$\boldsymbol{G}_l = \sum_{r=1}^{R_b} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r) \, d \, \frac{1}{\Sigma_r(l)} \boldsymbol{\xi}_r \, \boldsymbol{\xi}_r^{\top} \qquad \boldsymbol{K} = \sum_{r=1}^{R_p} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(r) \, \frac{1}{\sigma_r^2} \, \boldsymbol{\phi}_r \, \boldsymbol{\phi}_r^{\top},$$

- where $\Sigma_r(l)$ is the l-th diagonal element of the diagonal covariance matrix Σ_r, and $o_s(l)$ is the l-th element of the observation vector o_s.

Maximum Likelihood Linear Regression Based on Hidden Semi-Markov Model

- It is not always possible to estimate W and X for every distribution, because the amount of adaptation data of a target style is limited.
- Therefore, we use tree structures to group the distributions in the model and to tie the transformation matrices in each group in the same manner as HMM-based techniques.
- R_b and R_p is respectively the number of the distributions of the state output and duration distributions belonging to this group.

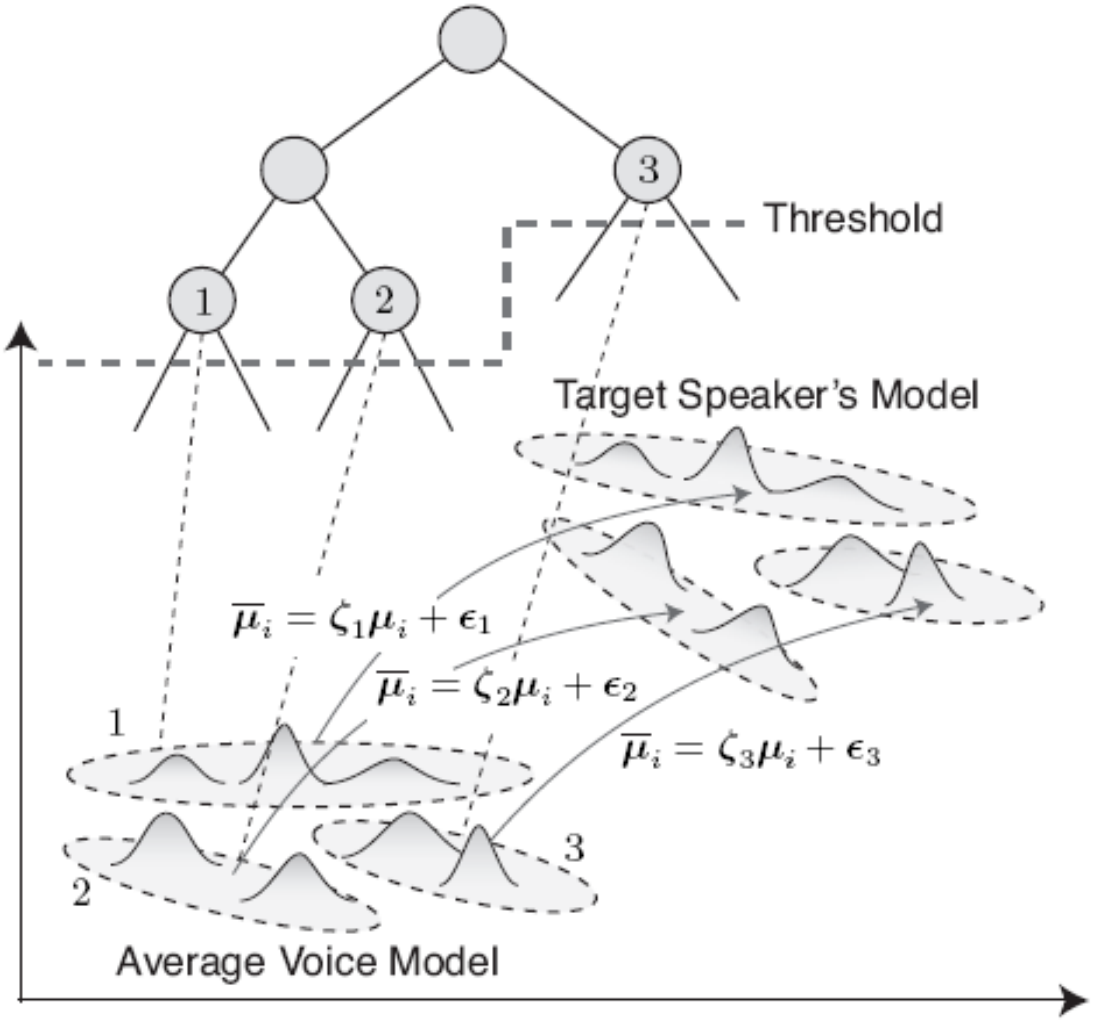


Figure 7.4: Maximum Likelihood Linear Regression

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 a male speaker MTK and a female speaker FTK as target speakers of the speaker adaptation and used the rest as training speakers for the average voice model.

Experimental Conditions

- In the system, gender-dependent average voice models were trained using 453 sentences for each training speaker.

- The total numbers of training sentences were 2265 sentences for male-speaker average voice model and 1812 sentences for female-speaker average voice model, respectively.

Objective Evaluation of the HSMM-Based MLLR Adaptation

- We calculated the target speakers' average mel-cepstral distance and root-mean-square (RMS) error of log F0 and vowel duration as the objective measure.
- The adaptation data was from 5 sentences to 450 sentences.
- Fifty test sentences were used for the evaluation, and these were included in neither the training nor the adaptation data.

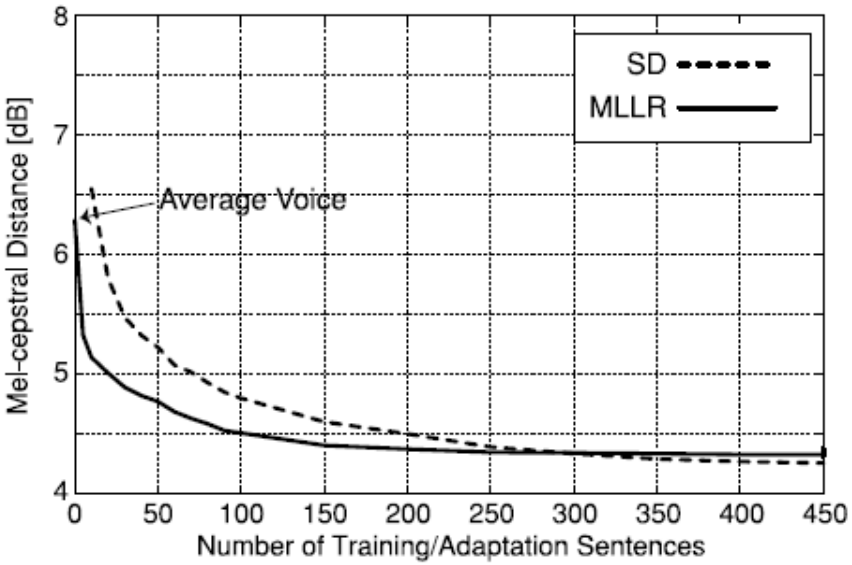


Fig. 11 Average mel-cepstral distance of male speaker MTK.

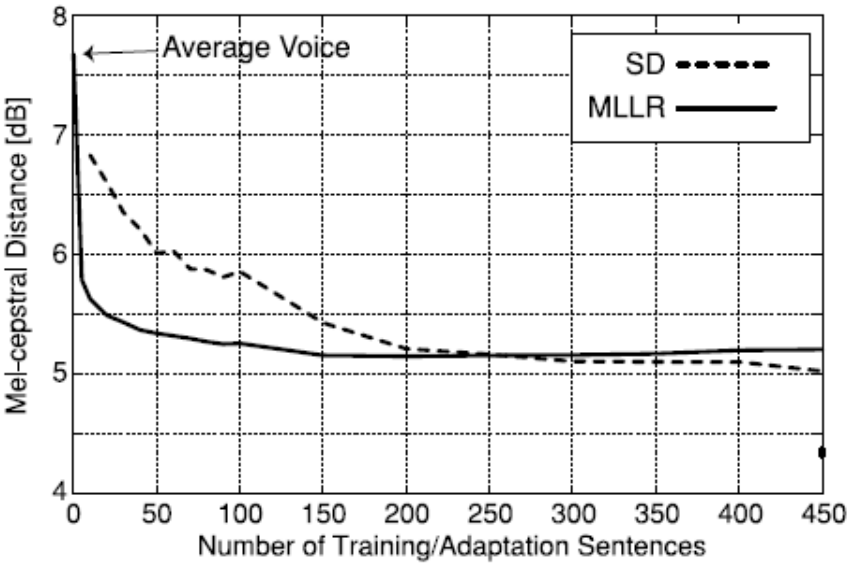


Fig. 14 Average mel-cepstral distance of female speaker FTK.

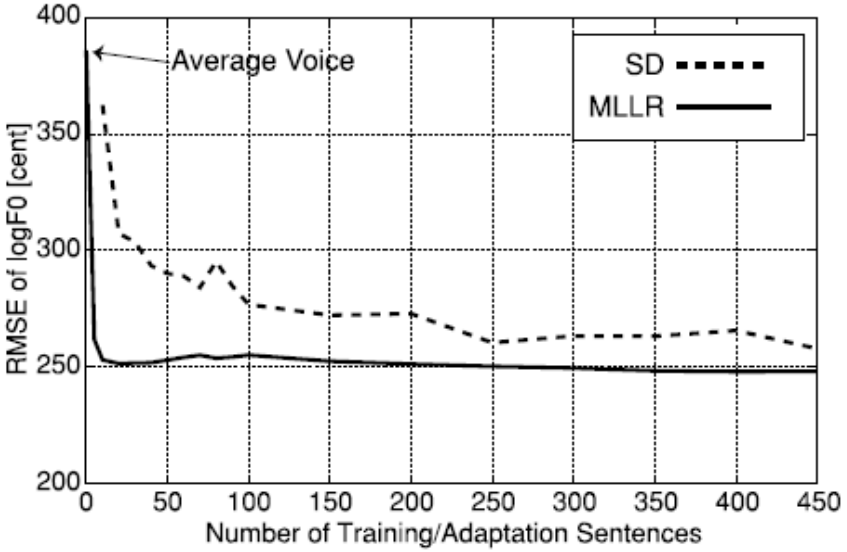


Fig. 12 RMS log F0 error of male speaker MTK.

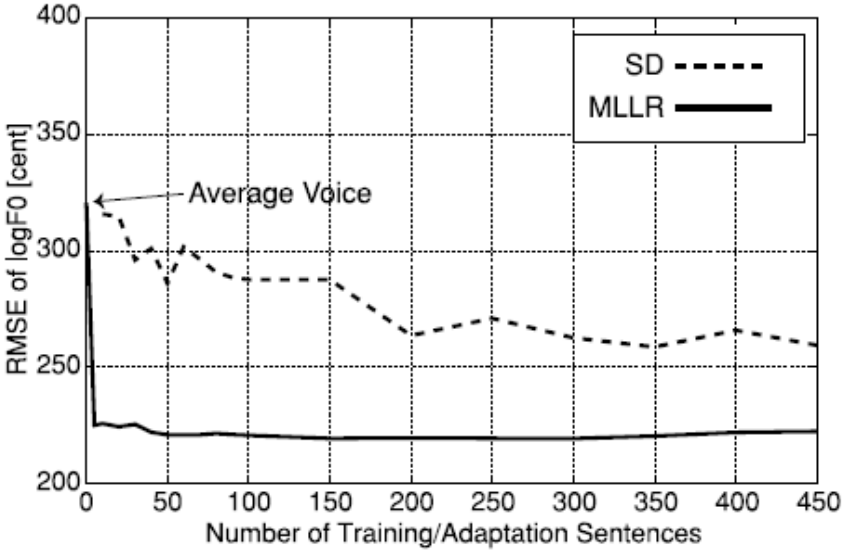


Fig. 15 RMS log F0 error of female speaker FTK.

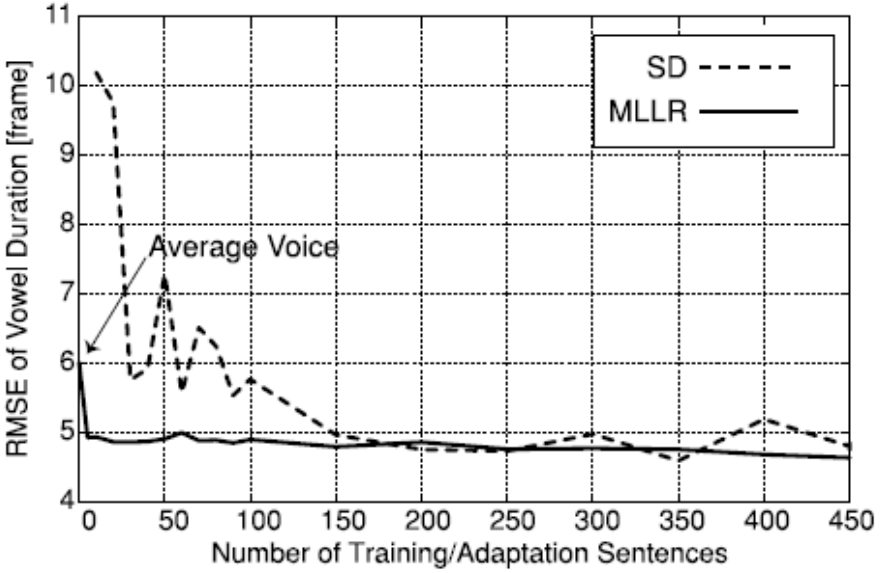


Fig. 13 RMS error of vowel duration of male speaker MTK.

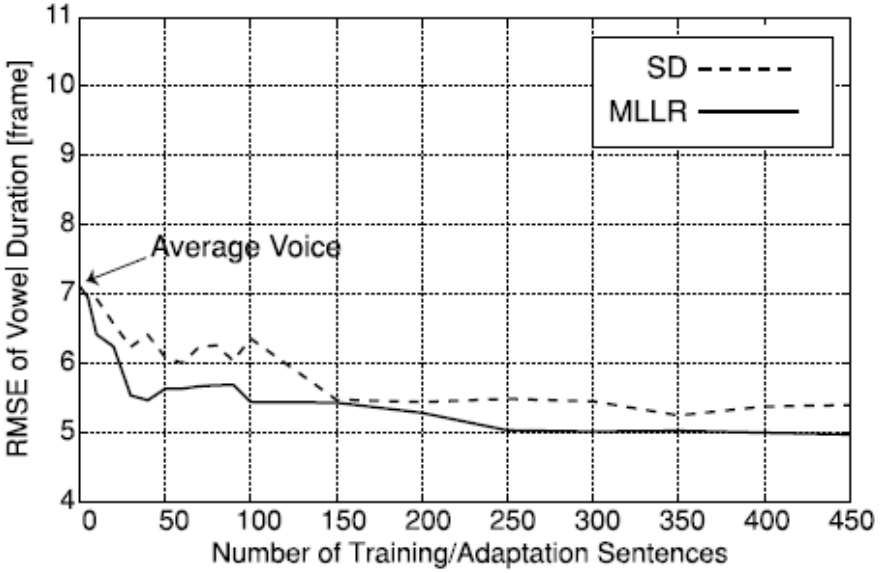


Fig. 16 RMS error of vowel duration of female speaker FTK.

Subjective Evaluation of the HSMM-Based MLLR Adaptation

- We compared the synthesized speech generated from the adapted models using 5, 10, 20, 50, 100, 150, 250, 350, and 450 sentences of the target speaker.

- For each subject, five test sentences were randomly chosen from a set of 50 test sentences, which were contained in neither the training nor the adaptation data.

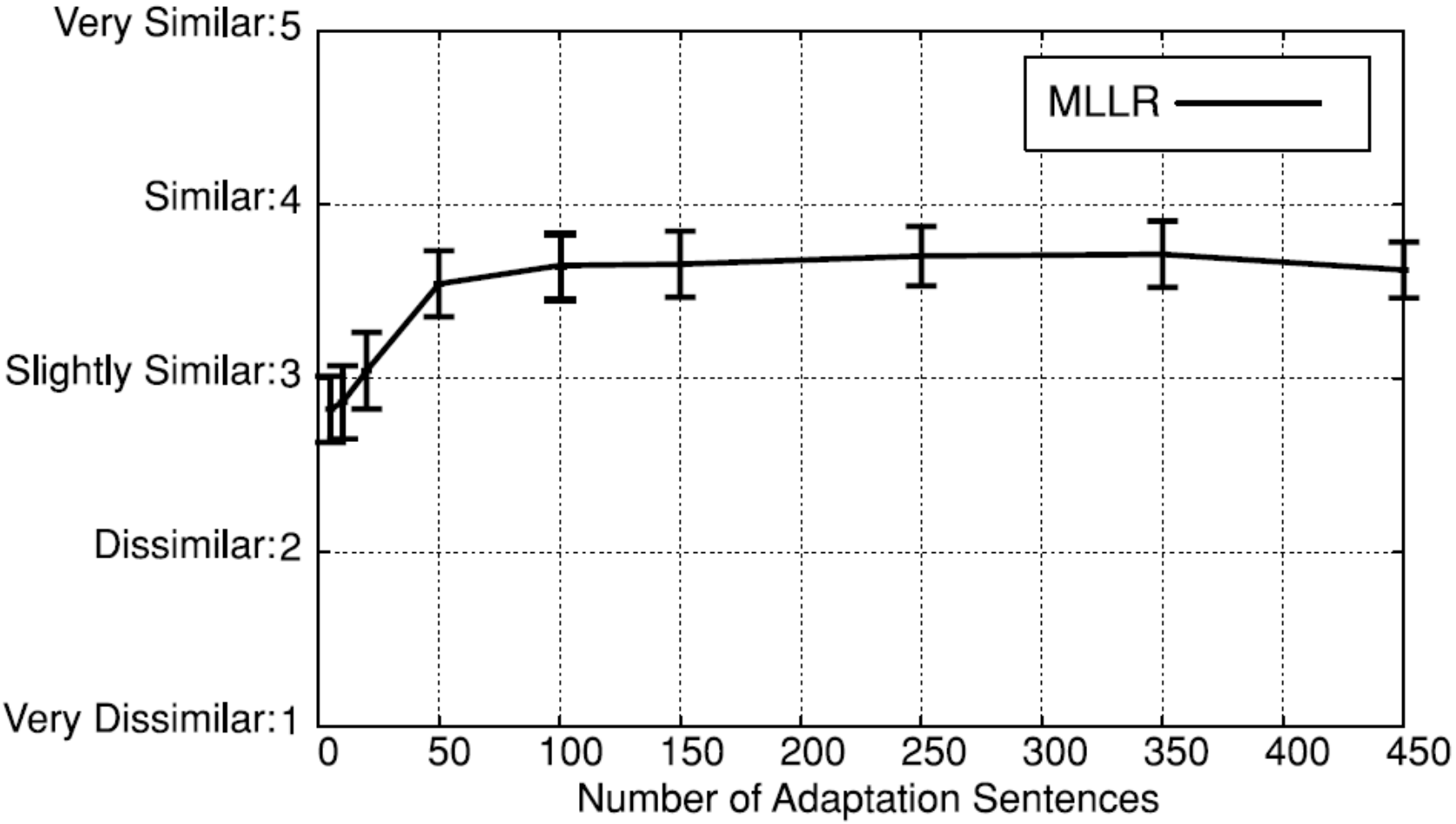


Fig. 19 Subjective Evaluation of the effect of the number of the adaptation data.

Subjective Evaluation of the HSMM-Based MLLR Adaptation

- We compared the synthesized speech generated from eight models with or without the adaptation of spectrum, F0, and/or duration.
- The adaptation data comprised 100 sentences.
- For reference, we also compared synthesized speech generated from the SD model using 453 sentences of the target speaker.

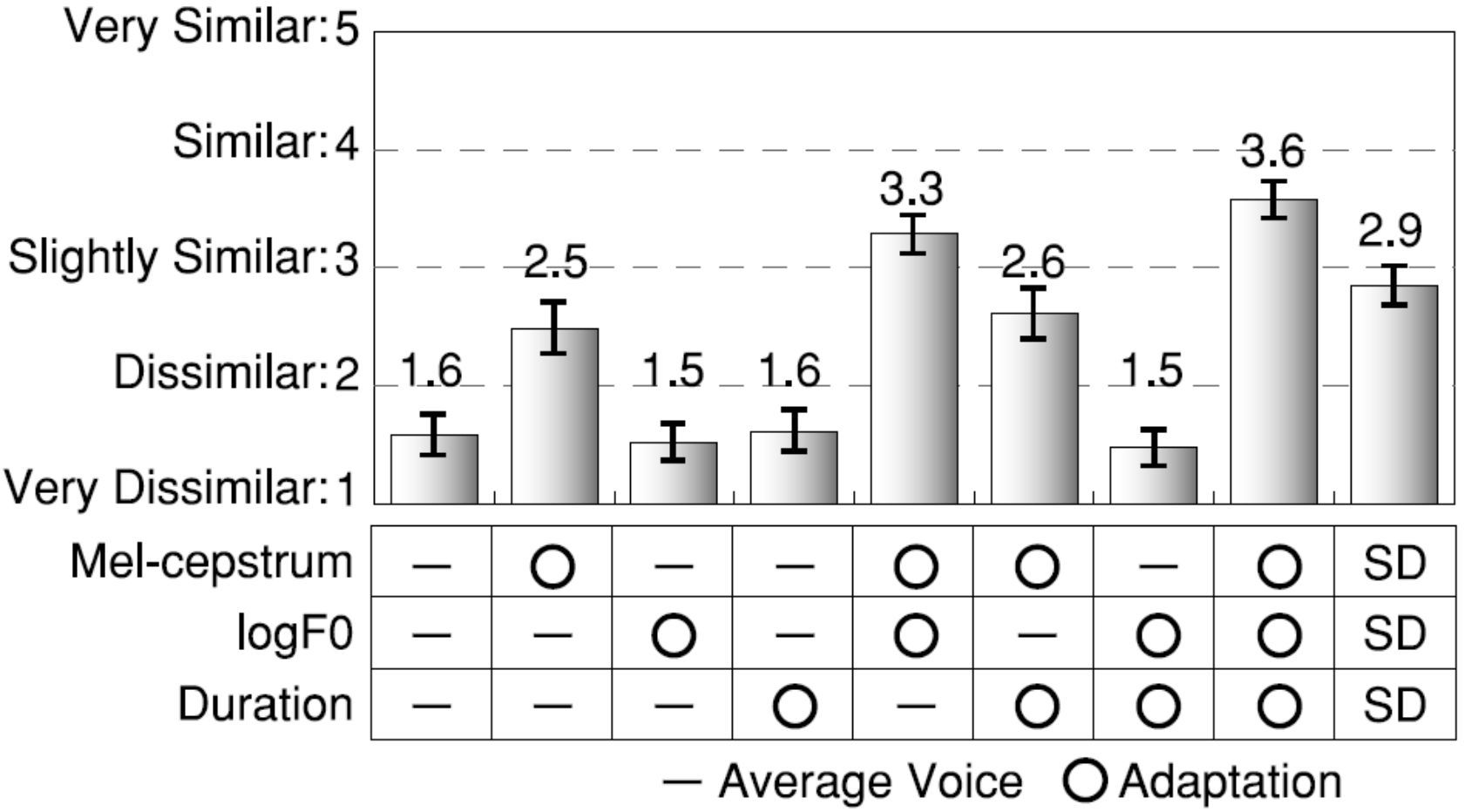


Fig. 20 Subjective evaluation of adaptation effects of each feature.