SPEAKER NORMALIZATION AND ADAPTATION BASED ON LINEAR TRANSFORMATION

Jun Ishii and Masahiro Tonomura

ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

ABSTRACT

We propose novel speaker independent (SI) modeling and speaker adaptation based on a linear transformation. An SI model and speaker dependent (SD) models are usually generated using the same preprocessing of acoustic data. This straightforward preprocessing causes a serious problem. Probability distributions of the SI models become broad and the SI models do not give good initial estimates for speaker adaptation. To solve these problems, a normalized SI model is generated by removing speaker characteristics using a shift vector obtained by the maximum likelihood linear regression (MLLR) technique. In addition, we propose a speaker adaptation method that combines the MLLR and maximum a posteriori (MAP) techniques from the normalized SI model. Experiments have been performed on Japanese phoneme recognition test using continuous density mixture Gaussian HMMs. For the baseline recognition test of normalized SI model, 12.8% reduction phoneme recognition error rate compared to the conventional SI model was achieved. Furthermore the proposed adaptation method using normalized SI model was effective than the tested conventional method regardless the amount of adaptation data.

1. INTRODUCTION

For practical use of speech recognition in many applications, speaker independent (SI) speech recognition systems using continuous mixture density HMMs (CDHMM) have been developed. SI models have many parameters that are trained using a large amount of data to cope with speech variations of many speakers. Performance of the SI model, however, is still poorer than that of a well trained speaker dependent (SD) model. Therefore, speaker adaptation is widely used to adapt the SI model to a specific speaker[1][2][3][4][5][6][7].

The SI models are usually constructed using various kinds of speaker independent speech with the same preprocessing for all speakers. This straightforward preprocessing causes a serious problem. The probability distributions of the SI models become broader than those of the SD model and the SI models do not give good initial estimates for speaker adaptation. To solve these problems, various speaker normalization techniques have been investigated[4][5][8][9][10].

This paper presents a novel SI modeling and speaker adaptation based on linear transformation. A normalized SI model is generated by removing speaker characteristics using a shift vector obtained by the maximum likelihood linear regression (MLLR) [1] technique. We also present a speaker adaptation method that combines the MLLR and maximum a posteriori (MAP) [3] techniques from the normalized SI model.

In the following section, we begin with an explanation of speaker normalization that uses linear transformation. Next, linear transformation-speaker adaptation using MAP estimation is described. In Section **3**, experimental results for a Japanese phrase database are given.

2. SPEAKER NORMALIZATION AND ADAPTATION USING LINEAR TRANSFORMATION

2.1. Linear transformation

In adaptation that uses linear transformation, the mean vector of a k-th Gaussian distribution to be adapted, $\hat{\mu}_k$, is calculated from the mean vector of the initial model μ_k:

$$\hat{\mu}_k = A\mu_k + b, \tag{1}$$

where A is an $n \times n$ (n is the order of the mean vector dimension) transformation matrix and b is a constant shift vector. In maximum likelihood linear regression adaptation (MLLR), which is one of the efficient linear transformation adaptation methods, A and b are estimated by maximizing the likelihood of the adaptation data.

Assuming that covariance matrices ($n \times n$) of all distributions are diagonal ($diag[\sigma^2_{k_1}, \sigma^2_{k_2}, \ldots, \sigma^2_{k_n}]$), the p-th row element of the transformation matrix A and the p-th element of b can be calculated by

$$\begin{bmatrix} a_{p,1} \\ \vdots \\ a_{p,n} \\ b_p \end{bmatrix} = \begin{bmatrix} g^{(p)}_{1,1} & \cdots & g^{(p)}_{1,n+1} \\ \vdots & & \vdots \\ g^{(p)}_{n,1} & & g^{(p)}_{n,n+1} \\ g^{(p)}_{n+1,1} & \cdots & g^{(p)}_{n+1,n+1} \end{bmatrix}^{-1} \begin{bmatrix} z^{(p)}_1 \\ \vdots \\ z^{(p)}_n \\ z^{(p)}_{n+1} \end{bmatrix}, \tag{2}$$

where

$$g^{(p)}_{i,j} = \sum_{k \in \Omega} \sum_{t=1}^{T} \gamma_k(t) \frac{\mu_{k_i} \mu_{k_j}}{\sigma^2_{k_p}} \tag{3}$$

$$(1 \le i, j \le n)$$

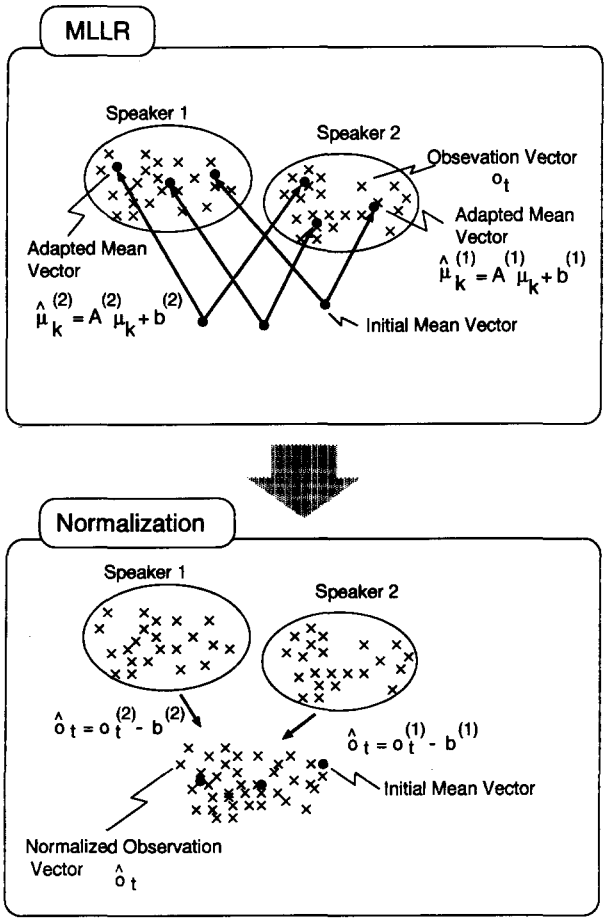

Figure 1. An example of normalizing observation vectors using MLLR.

$$g_{i,n+1}^{(p)} = g_{n+1,i}^{(p)} = \sum_{k\in\Omega}\sum_{t=1}^{T}\gamma_k(t)\frac{\mu_{k_i}}{\sigma_{k_p}^2} \quad (1 \le i \le n) \tag{4}$$

$$g_{n+1,n+1}^{(p)} = \sum_{k\in\Omega}\sum_{t=1}^{T}\gamma_k(t)\frac{1}{\sigma_{k_p}^2} \tag{5}$$

$$z_i^{(p)} = \sum_{k\in\Omega}\sum_{t=1}^{T}\gamma_k(t)\frac{\mathbf{o}_{t_p}\mu_{k_i}}{\sigma_{k_p}^2} \quad (1 \le i \le n) \tag{6}$$

$$z_{n+1}^{(p)} = \sum_{k\in\Omega}\sum_{t=1}^{T}\gamma_k(t)\frac{\mathbf{o}_{t_p}}{\sigma_{k_p}^2}, \tag{7}$$

and Ω denotes the shared Gaussian distribution set, $\gamma_k(t)$ denotes the posteriori probability of occupying the k-th Gaussian distribution at time t, and $\mathbf{o}_{t_p}$ is the p-th element of the observation vector. Here, μ_{k_i} and $\sigma_{k_p}^2$ are the i-th element of the mean vector and p-th element of variance, respectively.

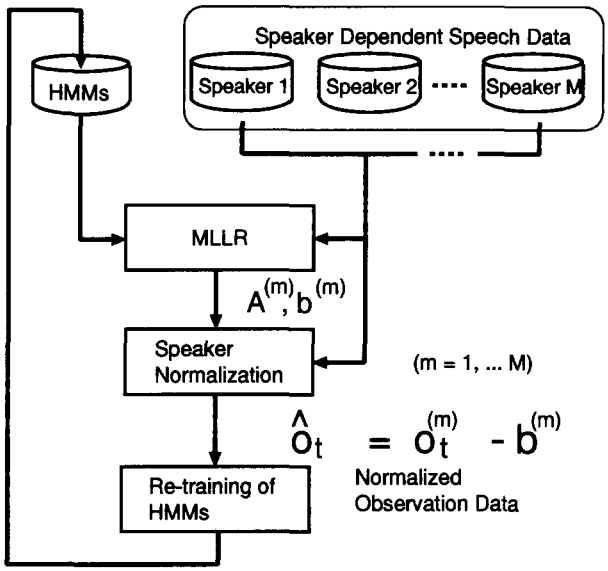

Figure 2. Procedure for generating a normalized speaker independent model.

2.2. Speaker normalization using MLLR shift vector

The transformation matrix A is regarded as a frequency warping representation, e.g. caused by vocal tract length difference. The constant shift vector b is considered to represent a specific speaker characteristic[2], implying that it can be used to remove speaker characteristic distribution. Then, normalized SI models can be constructed from normalized training data , in which shift vector b is subtracted from observation vector $\mathbf{o}_t$. Figure 1 shows an example of normalizing observation vectors using MLLR, and Figure 2 shows the procedure for generating a normalized SI model using the speech data of M-speakers. Letting $\mathbf{o}_t^{(m)}$ be the raw observation vector of the training data for speaker m, $A^{(m)}$ and $b^{(m)}$ denote the transformation coefficients of speaker m. The normalized SI model can be obtained with the following steps.

1. Estimate transformation coefficients ($A^{(m)}$ and $b^{(m)}$) for each speaker from training data using the current SI model.
2. Obtain the normalized observation vector $\hat{\mathbf{o}}_t$ by subtracting $b^{(m)}$:
$$\hat{\mathbf{o}}_t = \mathbf{o}_t^{(m)} - b^{(m)}. \tag{8}$$
3. Re-train the SI model parameters by using the normalized training data.
4. Iterate from Steps 1 to 3.

Note that the unnormalized SI model is used for the initial SI model in Step 1.

2.3. Speaker adaptation using MLLR and MAP techniques

For speaker adaptation from the normalized SI model described in **2.2**, we try to combine the existing speaker adaptation techniques: MLLR and MAP estimation. Both the

transformation matrix A and the shift vector b are estimated from adaptation data that is based on maximum likelihood criterion. Then, MAP estimates of the adapted mean vector for a k-th Gaussian distribution can be obtained by

$$\mu_k^{MAP} = A_k^{MAP}\mu_k + b_k^{MAP}, \tag{9}$$

where

$$A_k^{MAP} = \frac{\sum_{t=1}^{T}\gamma_k(t)A + I\tau_k}{\sum_{t=1}^{T}\gamma_k(t) + \tau_k} \tag{10}$$

$$b_k^{MAP} = \frac{\sum_{t=1}^{T}\gamma_k(t)b}{\sum_{t=1}^{T}\gamma_k(t) + \tau_k}, \tag{11}$$

and I denote the $n \times n$ identity matrix and τ_k indicates the weighting of *a priori* knowledge to empirical data.

Figure 3 shows an example of mean vector adaptation using MLLR with MAP estimation. In the figure, the thickness of a vector represents the total occupation probability of Gaussian distributions ($\sum_{t=1}^{T}\gamma_k(t)$). If the total occupation probability of a distribution is small, the adapted mean vectors using MLLR with MAP remain close to the initial mean vectors. On the other hand, if total occupation probability of a distribution is large, the mean vectors become close to the adapted mean vectors by using MLLR estimation. Thus, the mean adaptation is performed taking into consideration the reliability of MLLR estimation.

In the proposed method, we use a full matrix and individually estimate transformation matrices for each Gaussian distribution, while Digalakis et al. [11] and Zavaliagkos et al. [12] have proposed similar techniques. Digalakis use a diagonal matrix; in the paper of Zavaliagkos, the same matrix were used for shared distributions.

3. EXPERIMENTS

3.1. Conditions

The proposed algorithm was evaluated by using a Japanese 26-phoneme recognition test set. In the test, 279 phrases uttered by speakers that were not included in the SI model training were used.

The experimental conditions are listed in Table 1. The shared state triphone acoustic model (HMnet)[13] was used. The shared state structure of HMMs was determined by the SSS algorithm using 2620 isolated words uttered by one male speaker. The number of shared states was set to 200 with one additional state pause model. The number of mixture components per state was five. The SI model parameters were trained by using Baum–Welch algorithm with the 50 sentences uttered by 15 speakers, which were selected from among 285 speakers, and obtained by the speaker clustering method [14]. The shared Gaussian distributions set of MLLR estimation for normalization and adaptation (Ω in Section **2**) were constituted by all Gaussian distributions

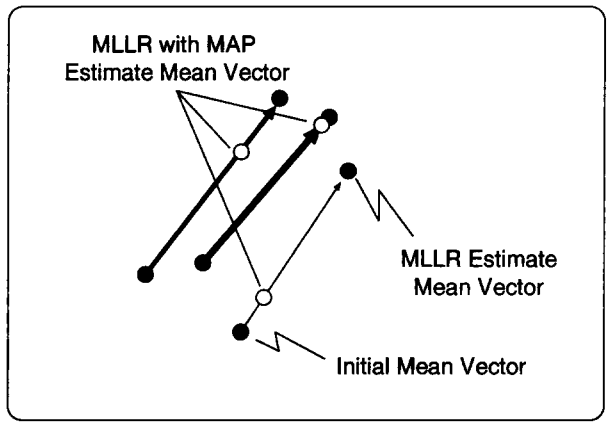

Figure 3. MLLR adaptation using MAP. Vector thickness represents total occupation probability of Gaussian distributions.

of HMMs. The training procedures are iteratively carried out three times.

Table 1. Experimental conditions.

Analysis conditions	
Sampling frequency: 12 kHz Hamming window: 20 ms; Frame period: 5 ms	
Analysis	
16-order LPC-Cepstrum + 16-order ΔLPC-Cepstrum + log power + Δlog power	
HMM	
200-state HMnet and 1-state HMM (pause) trained using 2620 words of one speaker	
Training data	
9-males + 6-females selected from 146 males + 139 females (50 Japanese sentences per person)	
Adaptation/Recognition data	
Speakers	3-males (MAU MMY MTM) 3-females (FAF FMS FYM)
Adaptation	N phrases from 598 Japanese phrases
Recognition	279 Japanese phrases

3.2. Results

Table 2 shows the effectiveness of the speaker normalization described in **2.2** for three male and three female speakers. Using the normalized SI model, the average baseline phoneme recognition error rate was reduced from 21.1% to 18.4% in comparison to the conventional SI model – the error reduction was 12.8%. was achieved. Performance of the proposed method was consistently better than that of the conventional method for each speakers. In particular, improved recognition performance was larger for speakers who showed lower performance using conventional SI model (FMS, FYM). This indicates that SI model normalization achieves a more accurate classification.

Table 2. Baseline recognition error rate comparison between the normalized SI model (upper) and the conventional SI model (lower) (%).

MAU	MMY	MTM	FAF	FMS	FYM	Total
15.2	15.2	12.0	20.2	18.4	29.5	18.4
15.5	17.0	13.3	21.9	25.2	33.4	21.1

Table 3. Recognition error rates of MLLR with MAP speaker adaptation from the normalized SI model (upper) and the conventional SI model (lower) (%).

speaker	# of adaptation phrases				
	3	5	7	10	20
MAU	15.8	15.0	14.9	15.2	13.7
	16.4	15.7	14.9	15.3	14.3
MMY	15.3	14.6	14.4	14.2	13.6
	17.3	16.0	16.0	15.3	14.6
MTM	11.8	11.8	11.0	10.9	9.9
	13.3	13.2	12.8	12.3	10.6
FAF	19.0	16.8	15.6	14.9	14.1
	21.8	19.8	18.5	16.5	15.1
FMS	19.5	18.5	17.7	16.6	13.9
	26.3	23.9	22.4	20.0	15.6
FYM	26.6	23.9	23.2	21.4	19.4
	29.6	24.0	25.4	24.2	19.6
Total	18.0	16.8	16.1	15.6	14.1
	20.8	18.8	18.3	17.2	14.9

Table 3 shows the phoneme recognition error rates of MLLR with MAP supervised adaptation (described in **2.3**) for the normalized SI model and for the conventional SI model while varying the number of adaptation phrases for six speakers. The adaptation training data was sampled from 598 phrases that were different from the test phrases. Considering the dependency on training data for the speaker adaptation performance, the experiment was repeated 3 times with different training data selections. τ of the MAP estimation in Section **2.3** was set to 4.0 for all distributions. As shown in the table, the performance of speaker adaptation for the normalized SI model was superior to that of the conventional SI model in all of the adaptation data. This proves that the normalized SI model has efficiently *a priori* knowledge. The proposed adaptation method makes good use of this *a priori* knowledge by using MAP estimation.

4. CONCLUSIONS

In this paper, novel SI modeling and speaker adaptation based on linear transformation techniques were described. In this method, a normalized SI model is generated by removing speaker characteristics using a shift vector obtained by the MLLR technique. In addition, we presented a speaker adaptation method that combines the MLLR and MAP techniques from the normalized SI model. Experimental results for a Japanese phoneme recognition test showed that the normalized SI model gave consistently better performance than the conventional SI model. The proposed adaptation method is more effective than the tested conventional method regardless the amount of adaptation data.

REFERENCES

[1] C. L. Leggetter and P. C. Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp. 171-185 (1995)

[2] S. J. Cox and J. S. Bridle: "Unsupervised speaker adaptation by probabilistic spectrum fitting," *Proc. of ICASSP 89*, pp. 294-297 (1989)

[3] C.-H. Lee, C.-H. Lin and B.-H. Juang: "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp. 806-814 (1991)

[4] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul: "Compact Model for Speaker-Adaptive Training," *Proc. of ICSLP 96*, pp. 1137-1140 (1996)

[5] N. Iwahashi: "Novel Training Method for Classifiers used in Speaker Adaptation," *Proc. of ICSLP 96*, pp. 2119-2122 (1996)

[6] M. Tonomura, T. Kosaka and S. Matsunaga: "Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum A Posteriori Probability Estimation," *Proc. of ICASSP 95*, pp. 688-691 (1995)

[7] J. Ishii, M. Tonomura, and S. Matsunaga: "Speaker Adaptation Using Tree Structured Shared-State HMMs," *Proc. of ICSLP 96*, pp. 1149-1152 (1996)

[8] Y. Ariki, S. Tagashira and M. Nishijima: "Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace," *Proc. of ICASSP 96*, pp. 319-322 (1996)

[9] E. Eide and H. Gish: "A Parametric Approach to Vocal Tract Length Normalization," *Proc. of ICASSP 96*, pp. 346-348 (1996)

[10] L. Lee and R. C. Rose: "Speaker Normalization Using Efficient Frequency Warping Procedures," *Proc. of ICASSP 96*, pp. 353-356 (1996)

[11] V. Digalakis and L. Neumeyer: "Speaker Adaptation Using Combined Transformation and Bayesian Method," *Proc. of ICASSP 95*, pp. 680-683 (1995)

[12] G. Zavaliagkos, R. Schwartz and J. McDonough: "Maximum a Posteriori Adaptation for Large Scale HMM Recognizers," *Proc. of ICASSP 96*, pp. 725-728 (1996)

[13] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," *Proc. of ICASSP 92*, pp. 573-576 (1992)

[14] T. Kosaka and S. Sagayama: "Tree-Structured Speaker Clustering For Fast Speaker Adaptation," *Proc. of ICASSP 94*, pp. 245-248 (1994)