SPEAKER INDEPENDENT ACOUSTIC MODELING USING SPEAKER NORMALIZATION

Jun Ishii [†‡] *and Toshiaki Fukada* [†]

[†] ATR Interpreting Telecommunications Research Labs.
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
[‡] MITSUBISHI Electric Corporation
5-1-1 Ohfuna, Kamakura, Kanagawa 247 Japan
ishii@media.isl.melco.co.jp

ABSTRACT

This paper proposes a novel speaker-independent (SI) modeling for spontaneous speech data from multiple speakers. The SI acoustic model parameters are estimated by individual training for inter-speaker variability and for intra-speaker phonetically related variation in order to obtain a more accurate acoustic model. The linear transformation technique is used for the speaker normalization to extract intra-speaker phonetically related variation and also is used for the re-estimation of inter-speaker variability. The proposed modeling is evaluated for a Japanese spontaneous speech data, using continuous density mixture Gaussian HMMs. Experimental results from the use of proposed acoustic model show that the reductions in word error rate can be achieved over the standard SI model regardless the type of acoustic model used.

1. INTRODUCTION

For practical use of speech recognition in many applications, speaker independent (SI) speech recognition systems using continuous mixture density HMMs (CDHMMs) have been developed. Recently, the speaker independent recognition of spontaneous speech has been studying.

The spontaneous speech data sampled from multiple speakers varies widely with regard not only to speech style but also context, training a SI model with this spontaneous speech data causes diffuseness and bias of HMM parameters. Furthermore, the number of speakers differs from one speech unit to another, and this difference results in the inter-speaker variability being represented incorrectly. As a consequence, the discriminating capability of the standard SI model is saturated, because of the indistinct estimation of HMM parameters is caused by mixing of inter-speaker variability factor and intra-speaker phonetically related variation factor.

These problems have been addressed by using speaker-normalization techniques, when generating the acoustic model. The signal bias removal technique[1][2] has been used to normalize the channel or speaker factor, and maximum likelihood linear regression (MLLR) [3], has been used to reduce inter-speaker variability[4][5][6][7]. Such speaker normalized (SN) models, however require channel or speaker adaptation prior to recognition process, because the models contain no inter-speaker acoustic variability. A recognizer without channel or speaker adaptation is desired for use in speaker independent speech recognition systems.

We therefor propose a method for producing acoustic model whose parameters are estimated by individual training for inter-speaker variability and for intra-speaker phonetically related variation. We call this named speaker-normalized and speaker-independent (SN-SI) modeling.

In the following section, we begin with an explanation of speaker normalized modeling that uses linear transformation. Next, generation of speaker normalization and speaker independent model is described. In Section 3, experimental results for a Japanese spontaneous database are given.

2. SPEAKER-NORMALIZED AND SPEAKER-INDEPENDENT MODEL

The speaker-normalized and speaker-independent (SN-SI) model is produced through the following two processing steps:

Step 1. A speaker normalized (SN) model is trained using speaker normalization based on a linear transformation technique. The HMM parameters are estimated by training for intra-speaker phonetically related variation.

Step 2. Inter-speaker variability is re-estimated by referring to the SN model to obtain the SN-SI model.

Figure 1 shows the conception of SN-SI modeling. Figure 1(b) expresses SN model distributions generated by the Step 1. in order to reduction of the diffuseness and bias from the standard SI model distribution (shown in Fig. 1(a)). The SN-SI model distributions shown in Fig. 1(c) represents inter-speaker variability with reference to the SN model.

2.1. Speaker-Normalized Model

2.1.1. Speaker Normalized Training Data

As shown in Figure 2, the SN model is generated based on speaker-normalization with linear transformation. The speaker normalization is performed by determining the set

$\bar{\mathbf{O}} = [\bar{O}^1 \bar{O}^2 \ldots \bar{O}^M]$ of speaker-normalized training data sequences from the set $\mathbf{O} = [O^1 O^2 \ldots O^M]$ of training data sequences obtained from the speech of M speakers (the observation sequence for speaker m is given by $O^m = [\mathbf{o}_1^m \; \mathbf{o}_2^m \; \ldots \; \mathbf{o}_{Tm}^m]$, where $\mathbf{o}$ is an nth-order vector and the subscript gives time). In this paper, the speaker adaptation technique is used to obtain the speaker normalized training data. We assume that the relative positions of the speaker adapted distributions and observation vectors give speaker normalized observation vectors.

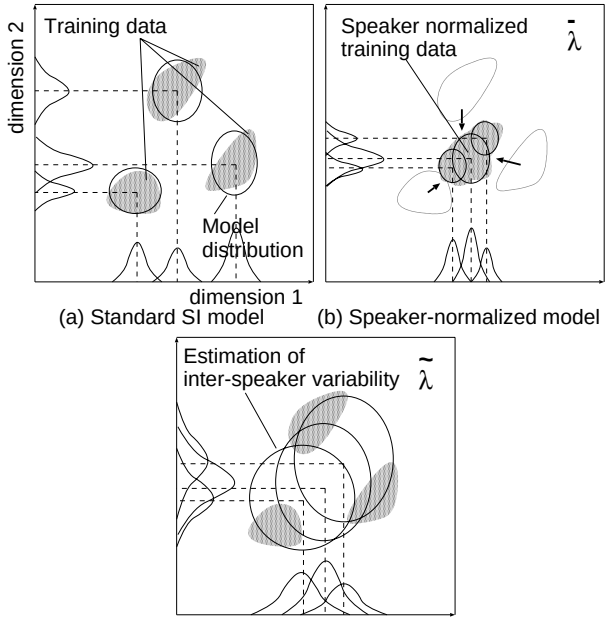

Figure 1: Output probability distribution of speaker normalized and speaker independent model.

The speaker adapted model $\hat{\lambda}^m$ for speaker m is obtained through speaker adaptation using the initial model λ and the training observation sequence O^m. The MLLR [3] is used for the speaker adaptation technique, and the mean vector $\mu_{j,k}$ (distribution k within state j) of Gaussian distribution is adapted to mean vector $\hat{\mu}_{j,k}$:

$$\hat{\mu}_{j,k}^m = A^m \mu_{j,k} + b^m, \tag{1}$$

where A^m is an $n \times n$ matrix and b^m is a nth-order vector, both of which are estimated for each shared class of Gaussian distributions.

The best state sequence $\mathbf{p}^m = [p_1^m p_2^m \ldots p_{Tm}^m]$ is then determined with the Viterbi algorithm by using $\hat{\lambda}^m$, O^m, and the context of the utterances. Furthermore, the Gaussian mixture distribution sequence $\mathbf{q}^m = [q_1^m q_2^m \ldots q_{Tm}^m]$, on which O^m indicates the maximum likelihood in every best state at each time, is extracted from the following equation:

$$q_t^m = arg \max_{q \in Q_t^m} \left[c_{p_t^m,q} \cdot \mathcal{N}(\mathbf{o}_t^m, \hat{\mu}_{p_t^m,q}^m, U_{p_t^m,q}) \right], \tag{2}$$

where Q_t^m is set of the mixture distributions within the best state sequence at time t, c is mixture weight, and U is the covariance matrix.

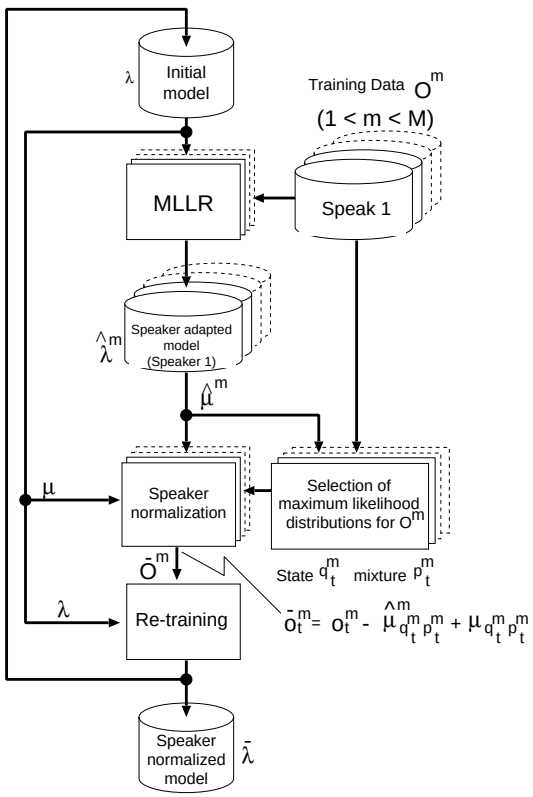

Figure 2: Procedure for generating a normalized speaker independent model.

The speaker-normalized observation sequence $\bar{O}^m = [\bar{\mathbf{o}}_1^m \; \bar{\mathbf{o}}_2^m \; \ldots \; \bar{\mathbf{o}}_{Tm}^m]$ is then obtained from the following equation using the mean vectors (before and after speaker adaptation) for the mixture distribution q_t in state p_t:

$$\bar{\mathbf{o}}_t^m = \mathbf{o}_t^m - \hat{\mu}_{p_t^m,q_t^m}^m + \mu_{p_t^m,q_t^m}, \tag{3}$$

The procedures described above are carried out for all the speakers in order to obtain set $\bar{\mathbf{O}}$ of speaker-normalized training data sequences.

2.1.2. Parameter Estimation of Speaker-Normalized Model

The speaker-normalized training data sequence $\bar{\mathbf{O}}$ is used for the re-training of λ. The HMM parameters are updated (mean vector $\bar{\mu}_{j,k}$, covariance matrix $\bar{U}_{j,k}$, mixture weight $\bar{c}_{j,k}$, transition probability $\bar{a}_{ij}$) with the following equations:

$$\bar{\mu}_{j,k} = \frac{\sum_{m=1}^{M} \sum_{t=1}^{T_m} \gamma_t^m(j,k) \cdot \bar{\mathbf{o}}_t^m}{\sum_{m=1}^{M} \sum_{t=1}^{T_m} \gamma_t^m(j,k)} \tag{4}$$

$$\bar{U}_{j,k} = \frac{\sum_{m=1}^{M}\sum_{t=1}^{T_m}\gamma_t^m(j,k)\cdot(\bar{\mathbf{o}}_t^m - \bar{\mu}_{j,k})(\bar{\mathbf{o}}_t^m - \bar{\mu}_{j,k})'}{\sum_{m=1}^{M}\sum_{t=1}^{T_m}\gamma_t^m(j,k)} \quad (5)$$

$$\bar{c}_{j,k} = \frac{\sum_{m=1}^{M}\sum_{t=1}^{T_m}\gamma_t^m(j,k)}{\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{T_m}\gamma_t^m(j,k)} \quad (6)$$

$$\bar{a}_{ij} = \frac{\sum_{m=1}^{M}\sum_{t=1}^{T_m-1}\xi_t^m(i,j)}{\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{t=1}^{T_m-1}\gamma_t^m(j,k)}, \quad (7)$$

where $\gamma_t^m(j,k)$ is the expected number of observations in state j mixture distribution k of $\bar{\mathbf{o}}_t^m$, $\xi_t^m(i,j)$ is the expected number of transitions from state i to state j, and K indicates the number of mixture in state j. The updated acoustic model replaces λ and the normalization is repeated several times. The final model obtained is called SN model $\bar{\lambda}$.

2.2. Generation of Speaker-Normalized and Speaker-Independent Model

The inter-speaker variability is expressed through the merging of Gaussian distributions after a speaker adapted model for each speaker is generated from the SN model.

1. As shown in Fig. 3(a), the adapted model for each target speakers is generated with MLLR using the SN model $\bar{\lambda}$ as the initial model.

$$\hat{\bar{\mu}}_{k,j}^m = \bar{A}^m\bar{\mu}_{j,k} + \bar{b}^m \quad (8)$$

2. Mean vector $\tilde{\mu}_{j,k}^m$ and covariance matrix $\tilde{U}_{j,k}$ are determined from equations (9) and (10) by using the mean vector $\hat{\bar{\mu}}_{j,k}^m$ of the adapted model and the covariance matrix $\bar{U}_{j,k}$ of the SN model to obtain the SN-SI model $\tilde{\lambda}$ (shown in Fig. 3 (b)), whose transition probability and coefficient of mixture weight are equal to those of SN the model.

$$\tilde{\mu}_{j,k} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\mu}}_{j,k}^m \quad (9)$$

$$\tilde{U}_{j,k} = \bar{U}_{j,k} + \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{\mu}}_{j,k}^m\cdot\hat{\bar{\mu}}_{j,k}^{m\prime} - \tilde{\mu}_{j,k}\cdot\tilde{\mu}_{j,k}' \quad (10)$$

3. EXPERIMENTS

3.1. Conditions

The proposed algorithm was evaluated by using Japanese spontaneous speech data set. The experimental conditions

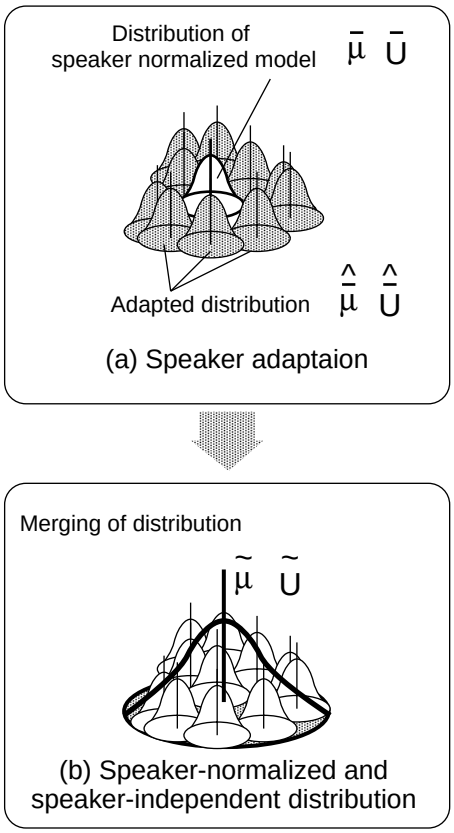

Figure 3: Generation of speaker normalized and speaker independent model.

are listed in Table 1. Travel arrangement conversation data recorded in ATR [8] was used for speech data. The male, female, and gender-independent acoustic models were generated using training data from 99 male speakers and 131 female speakers. The shared state structure of HMMs was determined using the ML-SSS (maximum likelihood successive state splitting)[9] algorithm, and a single state (10 mixtures) was used for the silent model. The MLLR for SN-SI modeling used 16 shared classes (determined by clustering the Gaussian distribution of initial SI model according to Kullback divergence). The vector b and diagonal component of regression matrix A were estimated. The normalization described in 2.1.2 was repeated, using the parameters estimated by the the Baum-Welch algorithm, five times. The variable-length N-gram[10] was used for the language model, and the recognition results were assessed with the first candidate from beam search[11] that utilized word graph.

3.2. Results

The continuous speech recognition test with the SN-SI model using the data of 16 male and 19 female speakers was performed. Table 2 shows the results of word error rates obtained when the number of HMM states was set to 401, 601, 801, and 1001. The table also shows the word error rates

Table 1: Experimental conditions.

Analysis	
• Sampling freq.	12k Hz
• Frame shift	10 ms
• Frame length	20 ms (Hamming window)
• Feature parameter	16-order cepstrum + 16-order Δcepstrum + log power + Δpower
Speech data	
• Travel arrangement task	
• Training	99 male speakers (13,000 words, 1237 utterances) 131 female speakers (20,000 words, 1725 utterances)
• Test	16 male speakers (2102 words, 196 utterances) 19 female speakers (2844 words, 244 utterances)
HMM	
• HMnet (5 mixtures/state) created with ML-SSS +silent model with 1 state (10 mixtures)	
Language model	
• Variable length N-gram – Training 414,326 words (6,396 different words) – Perplexity 19.34	

obtained with the SN model and the standard SI model.

The results from the use of SN-SI models, for all type of acoustic models, were superior to that of the standard SI model. In case of 601 states male model that gave biggest improvement in the test, the recognition error rate of SN-SI model was reduced from 46.8% to 42.6% in comparison to the standard SI model. This shows the effectiveness of individual training the parameters of inter-speaker variability and intra-speaker phonetically related variation. The discriminating capability was improved because of the reduction of the indistinct estimation of HMM parameters.

The SN model provided poorer recognition than dose either the SN-SI models or the standard SI models, because that contain no inter-speaker acoustic variability. To obtain the sufficient performance, the SN model requires speaker adaptation prior to recognition process.

4. CONCLUSIONS

This paper proposed a new acoustic modeling for spontaneous speech data from multiple speakers. To generate the accurate SI model, the HMM parameters were estimated by the individual for training inter-speaker variability and for intra-speaker phonetically related variation. This proposed modeling was evaluated for Japanese travel arrangement task using continuous density mixture Gaussian HMMs without adaptation, and the performances obtained using the proposed method was remarkable in comparison to the standard SI model regardless of the type of acoustic model.

5. REFERENCES

Table 2: Result of continuous word recognition – WER (%).
upper: **Speaker-normalized and speaker-independent model**
middle : Speaker normalized model
lower : Standard speaker-independent model

	Number of state of HMMs (Number of distributions)			
	401 (2010)	601 (3010)	801 (4010)	1001 (5010)
Male model (16 males)	**44.3**	**42.6**	**43.9**	**40.4**
	48.1	49.6	50.8	44.9
	45.4	46.8	45.7	41.2
Female model (19 females)	**29.9**	**28.7**	**30.3**	**28.6**
	32.9	32.1	35.0	35.0
	31.6	30.3	32.9	31.5
Gender-independent model (16 males, 19females)	**36.0**	**31.9**	**33.8**	**33.5**
	38.9	35.5	37.3	38.5
	37.7	33.8	34.7	35.4

[1] M. G. Rahim and B. H. Juang, "Signal Bias Removal for Robust Telephone Based Speech Recognition in Adverse Environments," *ICASSP94*, pp. 445-448, 1994.

[2] N. Iwahashi, "Novel Training Method for Classifiers used in Speaker Adaptation," *ICSLP96*, pp. 2119-2122, 1996.

[3] 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.

[4] T. Anastasakos, J. McDonoough, and J. Makhoul, "Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization," *ICASSP97*, pp. 1043-1046, 1997.

[5] D. Pye and P. C. Woodland, "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition," *ICASSP97*, pp. 1047-1050, 1997.

[6] J. McDonough T. Anastasakos, G. Zavaliagkos, and H. Gish, "Speaker-Adapted Training on the Switchboard Corpus," *ICASSP97*, pp. 1059-1062, 1997.

[7] J. Ishii and M. Tonomura: "Speaker Normalization and Adaptation Based on Linear Transformation," *ICASSP97*, pp. 1055-1058, 1997.

[8] A. Nakamura, S. Matsunaga, and T. Shimizu, "Japanese Speech Database for Robust Speech Recognition," *ICSLP96*, pp. 2199-2202, 1996.

[9] M. Ostendorf and H. Singer, "Maximum Likelihood Successive State Splitting," *Computer Speech and Language*, No. 11, pp. 17-41, 1997.

[10] H. Masataki and Y. Sagisaka, "Variable-Order N-Gram Generation by Word-Class Splitting and Consecutive Word Grouping," *ICASSP96*, pp. 188-191, 1996.

[11] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, and Y. Sagisaka, "Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graphs," *ICASSP96*, pp. 145-148, 1996.