TEXT-INDEPENDENT SPEAKER RECOGNITION BY COMBINING SPEAKER-SPECIFIC GMM WITH SPEAKER ADAPTED SYLLABLE-BASED HMM

Seiichi Nakagawa, Wei Zhang, Mitsuo Takahashi

Department of Information and Computer Sciences
Toyohashi University of Technology, Toyohashi, 441-8580, Japan

ABSTRACT

We presented a new text-independent speaker recognition method by combining speaker-specific Gaussian Mixture Model(GMM) with syllable-based HMM adapted by MLLR or MAP (EuroSpeech 2003[16]). The robustness of this speaker recognition method for speaking style's change was evaluated in this paper. The speaker identification experiment using NTT database which consists of sentences data uttered at three speed modes (normal, fast and slow) by 35 Japanese speakers(22 males and 13 females) on five sessions over ten months was conducted. Each speaker uttered only 5 training utterances (about 20 seconds in total). We obtained the accuracy of 98.8% for text-independent speaker identification for three speaking style modes (normal, fast, slow) by using a short test utterance (about 4 seconds). This result was superior to conventional methods for the same database. We show that the attractive result was brought from the compensational effect between speaker specific GMM and speaker adapted syllable based HMM.

1. INTRODUCTION

Speaker recognition has been a research topic for many years and various types of speaker models have been studied. Hidden Markov models (HMM) have become the most popular statistical tool for this task. The best results have been obtained using continuous density HMM (CHMM) for modeling the speaker characteristics [1]. For the text-independent task, where the temporal sequence modeling capability of the HMM is not required, one state CHMM, also called a Gaussian mixture model (GMM), has been widely used as a speaker model [2]. In accordance with [3], our previous study [4] showed that GMM can perform even better than CHMM with multi-states.

The objective of the speaker identification is to find a speaker model λ_i given the set of reference models $\Lambda = \{\lambda_1, \cdots, \lambda_N\}$ and sequence of test vectors (or frames) $X = \{x_1, \cdots, x_T\}$ which gives the maximum a posteriori probability $P(\lambda|X)$. This requires the calculation of all $P(\lambda_j|X), j = 1, \cdots, N$, and finding the maximum among them.

In most of the tasks, it is possible to use the likelihood $P(X|\lambda)$ instead of $P(\lambda|X)$ which does not to require prior probabilities $P(\lambda)$ to be known. Another simplifying assumption is that the sequence of vectors, X, are independent and identically distributed random variables. This allows to express $P(X|\lambda)$ as

$$P(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda), \tag{1}$$

where $P(x_t|\lambda)$ is the likelihood of single frame x_t given model λ. This is a fundamental equation of statistical theory and is widely used speech recognition. Generally speaking, $P(X|\lambda)$ is an utterance level score of X given model λ obtained from frame level scores $P(x_i|\lambda)$ using Eq. (1). Obviously, another ways of defining such scores can exist [5]. In GMM modeling techniques, feature vectors are assumed statistically independent, which is not true, but allows to simplify mathematical formulations. To overcome this assumption, recently, models based on segments of feature frames were proposed [6]. One of the disadvantages of GMM is that the acoustic variability dependent on phonetic events is not taken into account. Therefore, (large vocabulary continuous) speech recognition techniques have been used for text-dependent speaker identification [7]. The most attractive approach is to use a speaker adapted HMM from speaker-independent HMM [8]. This approach is also used for text-independent speaker identification. Sturim et al. used text-constrained GMM for text-independent speaker verification after segmenting input speech into pre-defined acoustic units by using speaker-independent speech recognizer [9]. Park et al. proposed a combination method of GMM and speaker-dependent segment-based speech recognizer [10]. The speaker-dependent speech recognizer is used for the segmentation results by a speaker-independent speech recognizer. Recently, Hazen et. al tried an integration method of GMM and speaker-dependent HMM [12]. In this paper, we propose a new combination method of speaker-specific GMM and speaker-adapted syllable-based HMM [16] and show the robustness.

2. SPEAKER MODELING

2.1. Gaussian Mixture Model (GMM)

A GMM is a weighted sum of M component densities and is given by the form

$$P(X|\lambda) = \sum_{i=1}^{M} c_i b_i(x), \tag{2}$$

where x is a d-dimensional random vector, $b_i(x), i = 1, \cdots, M$, is the component density and $c_i, i = 1, \cdots, M$, is the mixture weight. Each component density is a d-variate Gaussian function of the form

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\}, \tag{3}$$

with mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint that

$$\sum_{i=1}^{M} c_i = 1. \tag{4}$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from

all component densities. These parameters are collectively represented by the notation

$$\lambda = \{c_i, \mu_i, \Sigma_i\}, \qquad i = 1, \ldots, M. \tag{5}$$

In our speaker recognition system, each speaker is represented by such a GMM and is referred to by his/her model λ.

For a sequence of T test vectors $X = x_1, x_2, \cdots, x_T$, the standard approach is to calculate the GMM likelihood as in Eq. (1) which can be written in the log domain as

$$L(X|\lambda) = \log p(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda). \tag{6}$$

The GMM parameters are estimated by the E-M algorithm (the HTK toolkit [11]).

2.2. Speaker Adapted HMM

A parameter set of HMM is given by $\lambda = \{A, B, \pi\}$, where A, B and π denote a set of state transition probability, a set of output probability density functions, and a set of initial state probabilities, respectively. We used an acoustic model of a context-independent syllable-based HMM, which has a left-to-right topology and consists of 7states with 5 self-loops. Each output probability density function is represented by a 16 mixture Gaussian model with diagonal covariance matrices. The number of syllables is 124. Speaker adaptation is performed for B. The HMM parameters are estimated/adapted by the HTK toolkit. We describe in brief adaptation methods for a Gaussian distribution.

(i) MAP [11]

The speaker adaptation by Maximum A Posterior Probability Estimation (MAP) is in the following :

$$\hat{\mu}_N = \frac{(\alpha + N - 1)\hat{\mu}_{N-1} + X_N}{\alpha + N} = \frac{\alpha\mu_0 + \sum_{i=1}^{N} X_i}{\alpha + N}, \tag{7}$$

where $\{X_1, X_2, \cdots, X_m\}$ denotes training sample vectors and $N(\hat{\mu}_N, \hat{\Sigma}_N)$ denotes an estimated Gaussian Model adapted by training samples.

(ii) MLLR [11]

The speaker adaptation by Maximum Likelihood Linear Regression (MLLR) is defined as follows :

$$\hat{\mu} = A\mu_0 + b, \tag{8}$$

where A and b denote a regression matrix and an additive bias vector, respectively. These are estimated by using training samples.

3. SPEAKER IDENTIFICATION PROCEDURE

Figure 1 shows the structure of our speaker identification system. In this system, input speech is analyzed and transformed into a feature vector sequence by Front-end Analysis block and then each test vector x_t is fed to all reference speaker models of GMM and speaker adapted syllable-based HMMs in parallel. The i-th speaker dependent GMM produces likelihood $L^i_{GMM}(x), I = 1, 2, \cdots, N$. The i-th speaker adapted syllable-based HMMs also produce likelihood $L^i_{HMM}(x)$ by using a continuous syllable recognizer. All these likelihood values are passed in the so called likelihood decision block, where they are transformed to form the new score $L^i(x)$.

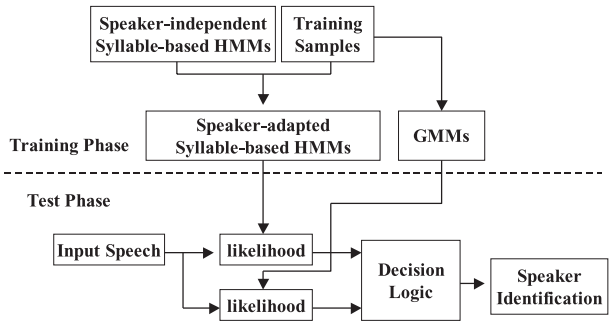

Fig. 1. Text-independent speaker identification by integration of GMM and speaker-adapted syllable-based HMMs

$$\mathcal{L}^i(X) = (1 - \alpha)L^i_{GMM}(X) + \alpha L^i_{HMM}(X), \tag{9}$$

where α denotes a weighting coefficient.

4. EXPERIMENTS

4.1. Database and Speech Analysis

For the experiments we used the NTT database.

The NTT database consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months (1990.8, 1990.9, 1990.12, 1991.3 and 1991.6) in a sound proof room [3]. For training the models, 5 same sentences for all speakers, from one session (1990.8) were used. Five other sentences uttered at normal, fast and slow speeds and same for each of the speakers, from the other four sessions were used as test data. Average duration of the sentences is about 4 sec. The input speech was sampled at 16KHz. 12 MFCC, their derivative (Δcep), and delta log-power were calculated at every 10ms with a window of 25 ms. Each session's mel-cepstrum vectors were mean normalized by mean subtraction and silence were removed.

4.2. Experimental Results

Figure 2 illustrates averaged text-independent speaker identification results on three different speaking styles (normal, fast, slow) by speaker-specific GMMs and speaker-adapted syllable-based HMMs for every utterance.

In the case of GMMs, we used 4 mixtures and 8 mixtures having full covariance matrices and 32 mixtures and 64 mixtures having diagonal covariance matrices, respectively. The parameters for three speaker-specific GMMs were estimated from only speaker-specific training data. We also conducted the estimation by using the MLLR based speaker adaptation method from the speaker-independent GMM (speaker-adapted GMM) [13]. The both were almost the same performance. When we integrate GMM with speaker-adapted HMM, however, we got better performance by speaker-specific GMMs than by speaker-adapted GMM. In the case of syllable-based HMMs, we obtained the likelihood from free syllable recognition without using any language models. The syllable recognition rate was about 80%. The rate was insensitive to speaker identification, because the identification rate was almost the same as the case of a text-dependent mode (i.e., syllable recognition rate = 100%). For GMMs, the GMM with 8-

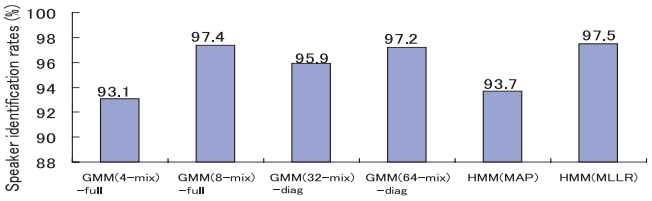

Fig. 2. Speaker identification (averaged) rates of text-independent speaker identification using normal, fast and slow speed test data for every utterance

mixtures was the best and it was comparable with the syllable-based HMM adapted by MLLR. For HMMs, the adapted HMMs by MLLR were better that those by MAP.

Figure 3 illustrates averaged text-independent speaker identification results by the combination of GMM with 8/64 mixtures and speaker adapted syllable-based HMMs by MLLR/MAP. We obtained the identification rate of 98.7% (99.4% for normal speed, 98.9% for fast speed and 97.9% for slow speed) in the combination case of GMM(8 mixtures) and HMM adapted by MLLR and 98.4% in the case of MAP, respectively. The combination improved the identification rate remarkably. On the other hand, Figure 4 illustrates the averaged results by combining two types of GMMs or HMMs adapted by MLLR and MAP for every utterance. The identification rates were about 98.1%(GMM) and 98.0%(HMM), which were worse than the rate in Fig.3 . It was caused by the fact that two types of GMMs or the speaker-adapted syllable-based HMMs by MLLR and MAP were similar to one another. On the other hand, the speaker-specific GMM and speaker-adapted syllable based HMM compensate their characteristics to one another.

Next, we investigated the identification performance for short test utterances. We took only two second segments from beginning parts in the above test utterances and identified the speaker. Figure 5 illustrates averaged text-independent speaker identification results by speaker-specific GMMs and speaker-adapted syllable-based HMMs. As expected, the performance became worse, that was 89.8% for GMM with 8 mixtures of full covariance matrices and 90.4% for HMM adapted by MLLR, respectively. As shown in Figure 6, however, we obtained the identification rate of 94.0%(96.0% for normal speed, 94.9% for fast speed and 91.0% for slow speed) in the combination case of GMM and MLLR-HMM and 93.2% in the case by GMM with 64 mixtures of diagonal covariance matrices.

Finally, we integrated two types GMMs(8 mixtures of full covariance matrices and 64 mixtures of diagonal matrices) with MLLR-HMM. The results for 2 seconds utterances are shown in Figure 7. Furthermore, the integration improved the identification (averaged) rate from 94.0% to 94.6%, especially, from 96.0% to 96.9% for normal speed. The integration of 3 methods improved the rate from 98.7% to 98.8% for every utterance.

5. DISCUSSIONS

The combination of speaker-specific GMM and speaker adapted syllable-based HMM improved the identification rate. So we investigated their compensational effect. Table 1 summarizes the 2 × 2 confusion matrix by the combination of GMM (8 mixtures)

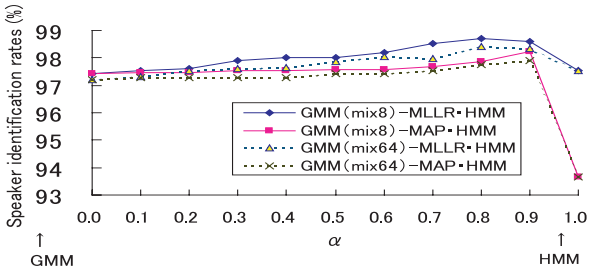

Fig. 3. Speaker identification (averaged) rates by integrating with text-independent speaker identification methods using normal, fast and slow speed test data for every utterance (GMM and MAP/MLLR-HMM)

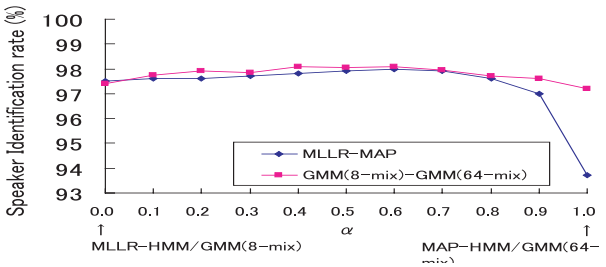

Fig. 4. Speaker-identification (averaged) rates by integrating with text-independent speaker identification methods for every utterance (GMM-8mix & GMM-64mix and MAP-HMM & MLLR-HMM)

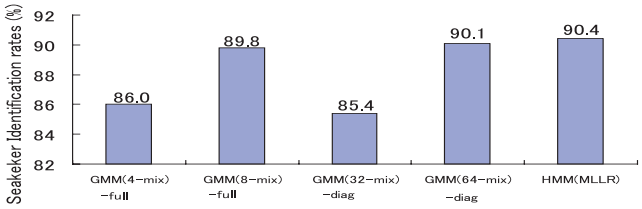

Fig. 5. Speaker identification (averaged) rates of text-independent speaker identification for every two seconds.

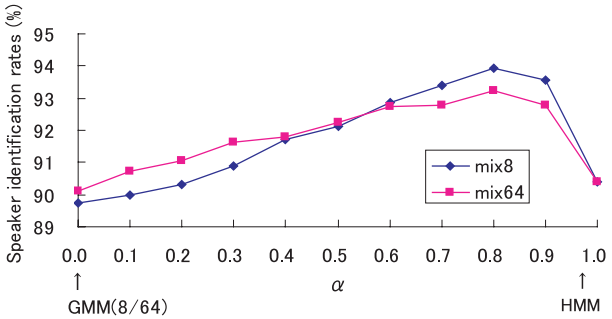

Fig. 6. Speaker identification (averaged) rates by integrating with text-independent speaker identification methods for every two seconds (GMM and MLLR-HMM).

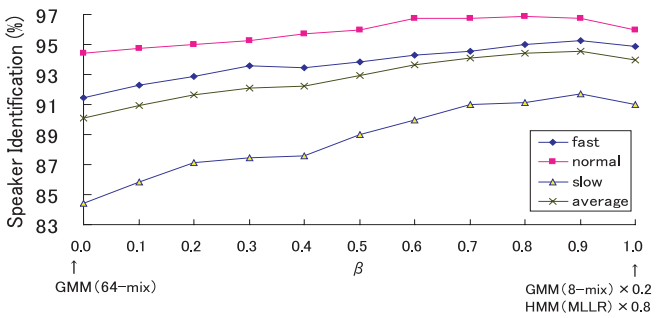

Fig. 7. Speaker identification rates by integrating with 3 methods(GMM-64 mixtures, GMM-8 mixtures, MLLR-HMM) for every two seconds

Table 1: Confusion matrix in Figure 3

GMM \ Syllable HMM	correct	incorrect
correct	2014 (2014)	32 (28)
incorrect	34 (27)	20 (4)

Table 2: An example of improvement (test speaker: Mito)

Speaker's Model	GMM likelihood	HMM likelihood	Combination likelihood
Mito	-66.62	-25.34	**-33.59**
Mkaw	**-66.55**	-25.53	-33.73
Mmik	-66.79	**-25.33**	-33.62
result	incorrect	incorrect	correct

and MLLR-HMM in Figure 3.

The total number of samples is 2100 (35 speakers × 4 sessions × 5 utterances × 3 speeds). The values in parentheses denote the number of correctly recognized samples after using the combination method. We can see the compensational effect. To our suprise, 4 samples out of 20 were correctly recognized, even if these samples were incorrectly recognized by both of GMM and syllable based HMM. We show a typical example in Table 2.

Our results were superior to the results by other studies for the same database[3,5,14,15,16]. For example, Miyajima et al. reported the rate of 99.0% for normal speaking rate utterances [14]. They used GMMs trained by 15 utterances, integrated by cepstrum coefficients and pitch and estimated by MCE.

6. CONCLUSION

We proposed a text-independent speaker recognition method by combining speaker specific GMM and speaker-adapted syllable-based HMM and we obtained the error reduction rate of about 50% for every utterance and 43% for every two seconds, respectively. From the speaker identification experiment using NTT database, we confirmed that our proposed method was superior to conventional text-independent speaker identification methods, showed the robustness and stated the reason.

7. REFERENCES

[1] Savic, M., Gupta, S., "Variable parameter speaker verification system based on Hidden Markov Modeling, in proceedings of ICASSP'90, pp. 281–284, 1990.

[2] Tseng, B., Soong, F., Rosenberg, A., "Continuous probabilistic acoustic map for speaker recognition", in proceedings of ICASSP'92, vol.II, pp. 161–164, 1992.

[3] Matusi, T., Furui, S., "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", in proceedings of ICASSP'92, vol.II, pp. 157–160, 1992.

[4] Markov, K., Nakagawa, S., "Text-independent speaker identification on TIMIT database", in proceedings of Acoustical Society of Japan, March 1995, pp. 83–94, 1995.

[5] Markov, K., Nakagawa, S., "Text-independent speaker recognition using non-linear frame likelihood transformation", Speech Communication, vol.24, pp.193–209, 1998.

[6] Liu, C.-S., Wang, H.-C., Soong, F. K., Huang, C.-S., "An orthogonal polynomial representation of speech signals and its probabilistic model for text independent speaker verification", in proceedings of ICASSP'95, vol.I, pp. 345–348, 1995.

[7] Matusi, T., Furui, S., "Concatenated phoneme models for text-variable speaker recognition", in proceedings of ICASSP'93, vol.II, pp. 391–394, 1995.

[8] Kanou, J., Katoh, M., Ito, A., Kohda, M., "A study on MLLR adpted speaker model for speaker verification", Technical Report on Spoken Language Processing, Infromation Processing Society of Japan, SLP29–10, 1999 (in Japanese).

[9] Sturim, D. E., et. al., "Speaker verification using text-constrained Gaussian", in proceedings of ICASSP2002, vol.I, pp. 677–680, 2002.

[10] Park, A., Hazen, T. J., "ASR dependent techniques for speaker identification", in proceedings of ICSLP2002, pp. 1337–1340, 2002.

[11] Young, S., Kershow,D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., "The HTK Book", 2000.

[12] Hazen, T.J., et. al., "Integration of speaker recognition into conversational spoken dialogue systems", Proc. EuroSpeech, pp.1961-1964, 2003.

[13] Reynolds, D.A, Quatieri, T.F., Dunn, R.B., "Speaker verfication using adapted Gaussian mixture models", Digital Signal Processing, vol.10, pp.19-41 (2000)

[14] Miyajima, C. Hattori, Tokuda. K., "Text-Independent Speaker Identification Using Gaussian Mixture Models Based on Multi-Space Probability Distribution", IEICE Trans, vol.E84-D, No.7, 847–855, 2001.

[15] Nishida, M. Ariki, Y, "Speaker recognition by separating phonetic space and speaker space", Proc. EuroSpeech, 1381–1384, 2001.

[16] Nakagawa,S.,Zhang, W., "Text-independent speaker recognition by speaker-specific GMM and speaker adapted syllable-based HMM", Proc. Eurospeech, pp.3017-3020 (2003)