

Text-independent Speaker Recognition by Speaker-specific GMM and Speaker Adapted Syllable-based HMM

Seiichi Nakagawa and Wei Zhang

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

{nakagawa, zhangwei}@slp.ics.tut.ac.jp

Abstract

We present a new text-independent speaker recognition method by combining speaker-specific Gaussian Mixture Model(GMM) with syllable-based HMM adapted by MLLR or MAP. The robustness of this speaker recognition method for speaking style's change was evaluated. 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. We obtained the accuracy of 100% for text-independent speaker identification. This result was superior to some conventional methods for the same database.

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 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 $\Lambda = \{\lambda_1, \dots, \lambda_N\}$ and sequence of test vectors (or frames) $X = \{x_1, \dots, x_T\}$ which gives the maximum a posteriori probability $P(\lambda|X)$. This requires the calculation of all $P(\lambda_j|X)$, $j = 1, \dots, N$, 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 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

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

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. Sturm 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. In this paper, we propose a new text-independent speaker recognition method by combining speaker-specific GMM with speaker-adapted syllable-based HMM.

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

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

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

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

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

$$\sum_{i=1}^M c_i = 1. \quad (4)$$

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

$$\lambda = \{c_i, \mu_i, \Sigma_i\}, \quad i = 1, \dots, M. \quad (5)$$

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 $X = x_1, x_2, \dots, x_T$, the standard approach is to calculate the GMM likelihood as in Eq. (1) which can be written in the log domain as

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

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 7 states with 5 self-loops. Each output probability density function is represented by a 16-mixed Gaussian model with diagonal covariance matrices. The number of syllables is 124 including syllables for loan words (Japanese consists of about 110 syllables). Speaker adaptation is performed for B . 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 :

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

$$\begin{aligned} \hat{\Sigma}_N &= \frac{1}{\beta + N} \{X_N X_N^T - (\alpha + N)\hat{\mu}_N \hat{\mu}_N^T \\ &\quad + (\beta + N - 1)\hat{\Sigma}_{N-1} \\ &\quad + (\alpha + N - 1)\hat{\mu}_{N-1} \hat{\mu}_{N-1}^T\}, \end{aligned} \quad (8)$$

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

(ii) MLLR [12]

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

$$\hat{\mu} = A\mu_0 + b, \quad (9)$$

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 a front-end analysis block and then each test vector x_t is fed to all reference speaker

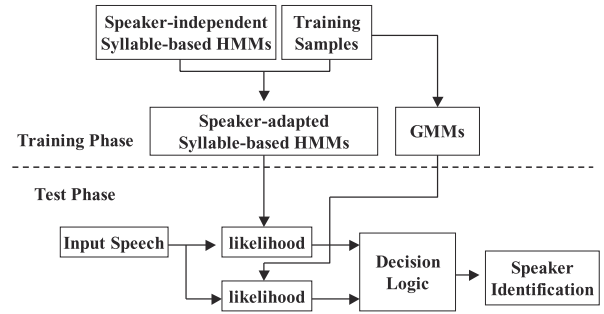


Figure 1: Text-independent speaker identification by integration of GMM and speaker-adapted syllable-based HMMs

models of GMM and speaker adapted syllable-based HMMs in parallel. The i -th speaker dependent GMM produces likelihood $L_{GMM}^i(x)$, $I = 1, 2, \dots, N$. The i -th speaker adapted syllable-based HMMs also produce likelihood $L_{HMM}^i(x)$ by using a continuous syllable recognizer. All these likelihoods are passed in the so called likelihood decision block, where they are transformed to form the new score $L^i(x)$.

$$\mathfrak{L}^i(X) = (1 - \alpha)L_{GMM}^i(X) + \alpha L_{HMM}^i(X), \quad (10)$$

where α denotes a weighting coefficient.

4. Experiments

4.1. Database and Speech Analysis

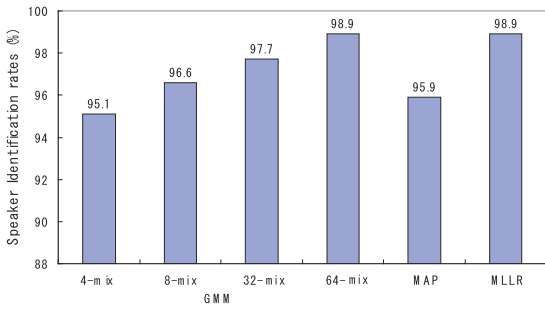
For the experiments we used the NTT database.

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 (Δ cep), 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 cepstrum mean subtraction and silence were removed.

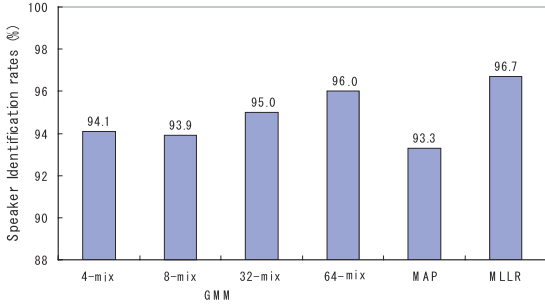
4.2. Experimental Results

Figure 2 illustrates text-independent speaker identification results by speaker-specific GMMs and speaker-adapted syllable-based HMMs.

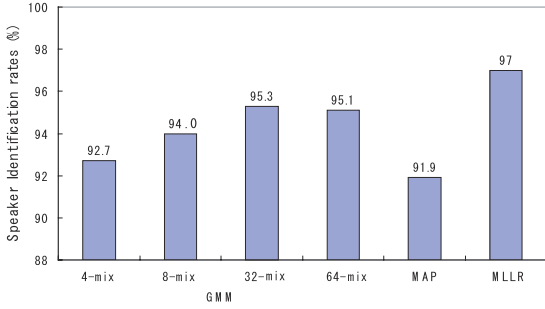
In the case of GMMs, we used 4 mixtures and 8 mixtures having full covariance matrices and 32 mixtures and 64 mixtures having diagonal covariance matrices, respectively. In the case of syllable-based HMMs, we obtained the likelihood from free syllable recognition without using any language models. The syllable recognition rate was about 80%. For GMMs, the GMM with 64-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 than those by MAP.



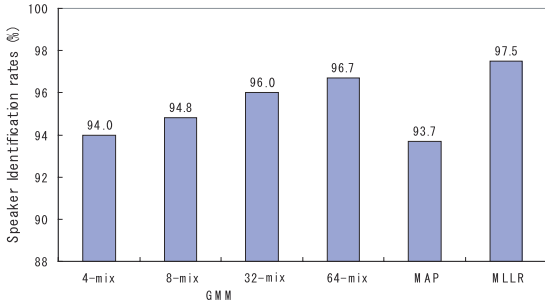
(a) Normal speed



(b) Fast speed

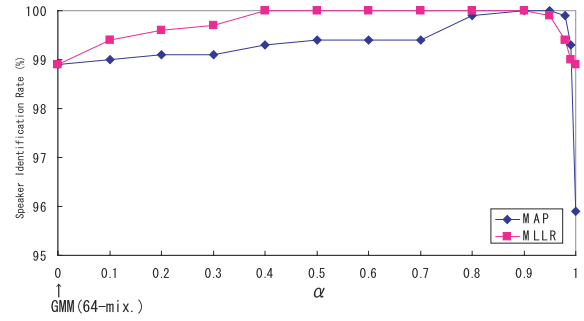


(c) Slow speed

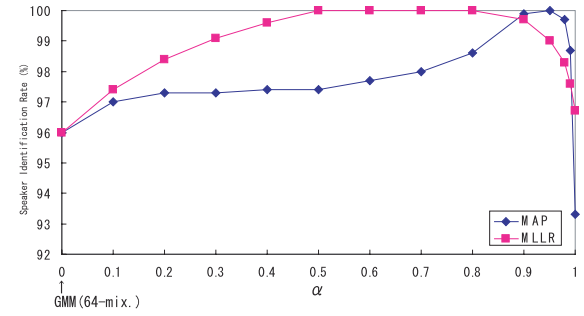


(d) average

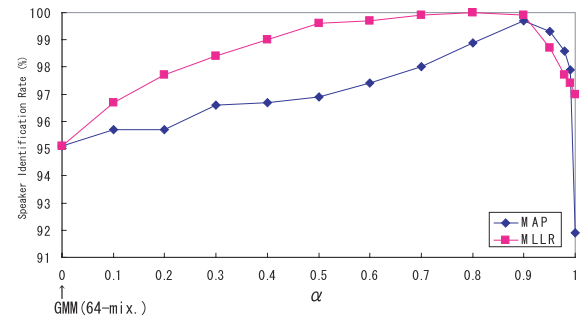
Figure 2: Speaker identification rates of text-independent speaker identification using normal, fast and slow speed test data



(a) Normal speed



(b) Fast speed



(c) Slow speed

Figure 3: Speaker identification rates by integrating with text-independent speaker identification methods using normal, fast and slow speed test data

Figure 3 illustrates text-independent speaker identification results by the combination of GMM with 64 mixtures and speaker adapted syllable-based HMMs. We obtained the identification rate of 100% for $\alpha=0.8$ in the case of MLLR and 99.9% for $\alpha=0.9$ in the case of MAP, respectively. The combination improved the identification rate remarkably.

Next, we investigated the identification performance for short test utterances. We took only two second segments from the above test utterances and identified the speaker. As expected, the performance became worse, that was, 89.2% for GMM and 90.4% for speaker-adapted HMM, respectively. However, we obtained the identification rate of 99.3% for $\alpha=0.7$ in the case of combination of GMM and MLLR (99.9% for normal speed, 99.4% for fast speed and 98.7% for slow speed).

Finally, we made speaker models by using only 3 training/adapting utterances. In this case, the 4-mixture Gaussian model with full covariance matrices was comparable with the 32-mixture Gaussian model with diagonal covariance matrices (94.0% on average). On the other hand, the performance by MLLR-based speaker-adapted HMM did not so degrade by small training samples (97.2% on average). By the combination of the 32-mixture Gaussian model and syllable-based HMM adapted by MLLR, we obtained the speaker identification of 100% for all speaking rate modes.

5. Related Work

We compare our method with related researches for the same NTT database; that is, for 22 males and 13 females. Matsui and Furui reported the speaker identification rate of 95.1%, 91.5% and 93.1% for normal, fast and slow speaking rate utterances, respectively. They used GMMs having 64 mixtures trained by 10 utterances[5].

Markov and Nakagawa reported the speaker identification rate of 97.3%, 93.4% and 93.0% for normal, fast and slow speaking rate utterances, respectively. They used GMMs having 8 mixtures with full covariance matrices trained by 10 utterances and non-linear frame likelihood transformation[3].

Miyajima et al. reported the rate of 99.0% for normal speaking rate utterances [13]. They used GMMs trained by 15 utterances, integrated by cepstrum coefficients and pitch and estimated by MCE.

Nishida and Ariki reported the rate of 94.9% for normal speaking rate utterances[14]. They used a subspace method, which maps separately speech to a phonetic space and a speaker characteristic space. The speaker model was trained by 5 utterances. Our proposed method outperformed the above related researches.

6. Conclusion

We proposed a text-independent speaker recognition method by combining speaker specific GMM and speaker-adapted syllable-based HMM. From the speaker identification experiment using NTT database, we confirmed that our proposed method was superior to conventional text-independent speaker identification methods. In near future, we will compare a speaker-specific GMM trained by only the speaker's utterances with a speaker-adapted GMM adapted from a speaker-independent HMM.

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 adapted speaker model for speaker verification", *Technical Report on Spoken Language Processing*, Information 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] Tsurumi, Y., Nakagawa, S., "An supervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation", in proceedings of ICSLP94, pp. 431–434, 1994.
- [12] Leggetter, C. J., Woodland, P. C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol.9, 171–185, 1995.
- [13] 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.
- [14] Nishida, M. Ariki, Y., "Speaker recognition by separating phonetic space and speaker space", *Proc. EuroSpeech*, 1381–1384, 2001.