IEICE TRANS. INF. & SYST., VOL.E91–D, NO.10 OCTOBER 2008

LETTER

Text-Independent Speaker Verification Using Artificially Generated GMMs for Cohorts

Yuuji MUKAI[†], *Nonmember*, **Hideki NODA**[†a)], **Michiharu NIIMI**[†], *and* **Takashi OSANAI**[††], *Members*

SUMMARY This paper presents a text-independent speaker verification method using Gaussian mixture models (GMMs), where only utterances of enrolled speakers are required. Artificial cohorts are used instead of those from speaker databases, and GMMs for artificial cohorts are generated by changing model parameters of the GMM for a claimed speaker. Equal error rates by the proposed method are about 60% less than those by a conventional method which also uses only utterances of enrolled speakers.

key words: *speaker verification, Gaussian mixture model, artificial cohort, score normalization, background model*

1. Introduction

Speaker verification (SV) is the task of determining whether the claimed identity of a speaker is correct. Statistical approaches using Gaussian mixture models (GMMs) are commonly used for text-independent SV [1]. An important issue in the statistical approaches is that of score normalization, and several normalization methods have already been proposed. Popular normalization methods, which are briefly reviewed in 2, are as follows: normalization using universal background model [1], cohort normalization method [2], and T-norm [3].

All of the above score normalization methods need speaker databases. However, preparation of a database is a burden particularly in a small scale SV system such as one for home security, where the number of enrolled speakers is very small*. In such an SV system, it is desirable to perform SV using only utterances of enrolled speakers, if it is possible. In fact, this issue is already addressed in text-dependent SV using hidden Markov model [4] and text-independent SV using GMM [5], though their purpose is to build a flexible and portable SV system running on portable devices such as palm-top computers and wireless phones.

In [5], a background model was estimated using the training data for a claimed speaker, i.e., the same data was used to build the claimed speaker model and its background model. The difference between the two models is the number of Gaussian mixtures, and 32-mixture GMM and 16-mixture GMM were used for the claimed speaker model and its background model, respectively. However, SV performance by this approach seems to be poor according to our experiments described in 4. This paper proposes an alternative approach for GMM-based text-independent SV using only utterances of enrolled speakers. We use artificial cohorts instead of those from speaker databases. Considering that GMMs for cohorts for a claimed speaker are relatively close to that for the claimed speaker, we generate GMMs for cohorts by changing model parameters of the GMM for the claimed speaker.

2. Score Normalization Methods for Speaker Verification

Let $\mathbf{Y} = \{\mathbf{y}_t; t = 1, \cdots, T\}$ denote a sequence of feature vectors obtained from input speech, and let $p_s(\mathbf{y}_t)$ and $p_o(\mathbf{y}_t)$ be probability density functions (pdfs) of $\mathbf{y}_t$ for a claimed speaker (true speaker) and all other possible speakers (impostors), respectively. Here both pdfs are modeled by GMMs. Assuming that $\mathbf{y}_t$s are mutually independent and then $p_s(\mathbf{Y}) = \prod_{t=1}^{T} p_s(\mathbf{y}_t)$ and $p_o(\mathbf{Y}) = \prod_{t=1}^{T} p_o(\mathbf{y}_t)$, log-likelihood ratio $S(\mathbf{Y})$ is given as

$$S(\mathbf{Y}) = \log \frac{p_s(\mathbf{Y})}{p_o(\mathbf{Y})} \tag{1}$$

$$= \sum_{t=1}^{T} \log \frac{p_s(\mathbf{y}_t)}{p_o(\mathbf{y}_t)}. \tag{2}$$

In fact, instead of $p_s(\mathbf{Y})$ and $p_o(\mathbf{Y})$, the normalized likelihood by the length T of the vector sequence $\mathbf{Y}$ (the number of frames), i.e., $p_s(\mathbf{Y})^{1/T}$ and $p_o(\mathbf{Y})^{1/T}$ are usually used. In that case, the log-likelihood ratio $S(\mathbf{Y})$ is given as

$$S(\mathbf{Y}) = \frac{1}{T} \log \frac{p_s(\mathbf{Y})}{p_o(\mathbf{Y})} \tag{3}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \log \frac{p_s(\mathbf{y}_t)}{p_o(\mathbf{y}_t)}. \tag{4}$$

Using S($\mathbf{Y}$), the decision on the hypothesis that $\mathbf{Y}$ is from

Manuscript received April 10, 2008.
Manuscript revised June 14, 2008.

[†]The authors are with the Department of Systems Design and Informatics, Kyushu Institute of Technology, Iizuka-shi, 820–8502 Japan.
[††]The author is with National Research Institute of Police Science, Kashiwa-shi, 277–0882 Japan.
a) E-mail: noda@mip.ces.kyutech.ac.jp

DOI: 10.1093/ietisy/e91–d.10.2536

*In general, speaking and recording conditions influence SV performance and in particular, handset (microphone) variability causes significant performance degradation in SV systems [1]. If speaking and recording conditions in such an SV system are similar to those under which speech data from many speakers have already been collected for a database, score normalization can be carried out using the universal background model or cohorts from the database. However in general, we cannot expect this coincidence on speaking and recording conditions. Therefore, preparation of a database is almost always required for any SV system.

the claimed speaker is made as follows.

$$S(\mathbf{Y}) \geq \theta, \quad \text{accept the hypothesis} \tag{5}$$

$$S(\mathbf{Y}) < \theta, \quad \text{reject the hypothesis,} \tag{6}$$

where θ is a decision threshold.

Taking the aforementioned general procedure into account, we review popular score normalization methods: normalization using universal background model [1], cohort normalization method [2], and T-norm [3].

(1) Use of universal background model Speech samples from a large number of speakers are used to train a single GMM for $p_o(\mathbf{y}_t)$, which is called a universal background model or a world model.

(2) Cohort normalization Cohort normalization uses a set of other speakers called cohorts whose pdfs are close to that for a claimed speaker. Cohorts for the claimed speaker are selected from speaker databases. Given the selected cohorts $c_i, i = 1, \cdots, N$ and their pdfs $p_{c_i}(\mathbf{Y})$, the following $p_o(\mathbf{Y})$,

$$p_o(\mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} p_{c_i}(\mathbf{Y}) \tag{7}$$

is used instead of the universal background model.

(3) T-norm T-norm extends the standard cohort normalization. The log-likelihood for the claimed speaker $\log p_s(\mathbf{Y})$ is normalized using the mean μ and standard deviation σ of the log-likelihood for cohorts $\log p_{c_i}(\mathbf{Y}), i = 1, \cdots, N$:

$$S_T = \frac{\log p_s(\mathbf{Y}) - \mu}{\sigma}. \tag{8}$$

The advantage of T-norm over the cohort normalization is the use of variance parameter σ which approximates the distribution of cohorts more accurately.

3. Artificial Cohort Model

After reviewing GMM, we describe how to generate GMMs for artificial cohorts.

3.1 Gaussian Mixture Model

A mixture of K Gaussian distributions is described as

$$p(\mathbf{y}_t) = \sum_{k=1}^{K} a_k g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k), \quad \sum_{k=1}^{K} a_k = 1, \tag{9}$$

$$g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{D/2} |\mathbf{\Sigma}_k|^{1/2}} \cdot \exp\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{m}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{y}_t - \mathbf{m}_k)\}, \tag{10}$$

where $\mathbf{y}_t$ is a D dimensional feature vector at t-th frame and a_k is the mixing coefficient of the k-th Gaussian distribution $g_k(\mathbf{y}_t; \mathbf{m}_k, \mathbf{\Sigma}_k)$ with mean vector $\mathbf{m}_k$ and covariance matrix $\mathbf{\Sigma}_k$. The model parameters, $a_k, \mathbf{m}_k, \mathbf{\Sigma}_k, k = 1, \cdots, K$ are iteratively estimated by the EM method [6]. Explicit procedures of the EM method are found in [7], [8]. The initial values to start the iterative procedure are obtained by clustering training samples using the VQ method [9].

3.2 GMMs for Artificial Cohorts

Considering that the pdfs for cohorts for a claimed speaker are relatively close to that for the claimed speaker, it could be possible to generate them artificially by modifying the pdf for the claimed speaker. We make GMMs for artificial cohorts by changing model parameters of the GMM for the claimed speaker.

Given model parameters, $a_k, \mathbf{m}_k, \mathbf{\Sigma}_k, k = 1, \cdots, K$ of the GMM $p_s(\mathbf{y}_t)$ for the claimed speaker, those parameters $a_k^{c_i}, \mathbf{m}_k^{c_i}, \mathbf{\Sigma}_k^{c_i}$ for artificial cohorts $c_i, i = 1, \cdots, N$ are here set as

$$a_k^{c_i} = a_k, \tag{11}$$

$$\mathbf{m}_k^{c_i} = \mathbf{m}_k + \alpha \mathbf{r}_k^{c_i}, \tag{12}$$

$$\mathbf{\Sigma}_k^{c_i} = \beta \mathbf{\Sigma}_k, \tag{13}$$

where α and β are parameters which should be set experimentally, and $\mathbf{r}_k^{c_i}$ is a random vector whose components $r_{k,d}^{c_i}, d = 1, \cdots, D$ are uniformly distributed in the interval $-1 \leq r_{k,d}^{c_i} \leq 1$. The parameter α controls variations of $\mathbf{m}_k^{c_i}$ for cohorts from $\mathbf{m}_k$, and the parameter $\beta > 1$ increases variances for cohorts from those for the claimed speaker. Appropriate values of α and β are described in 4.

4. Speaker Verification Experiments

For SV experiments, telephone speech data-set was used, which consists of isolated uttered Japanese 20 words produced two repetitions by 100 male speakers in two sessions spaced three to four months apart [10]. The speech data was low-pass filtered at 4.5 kHz and digitized at 10 kHz sampling rate. The digitized speech was pre-emphasized with a first-order adaptive filter and subjected to 12th order LPC analysis with 25.6 msec Hamming window and 12.8 msec frame rate. In fact, the selective LPC analysis was applied to use the spectral information up to 4 kHz considering that the speech data is telephone speech. The twelve LPC cepstral coefficients obtained by this analysis were used as a feature vector for each time frame.

The data-set was divided into two sets: one set consists of first-uttered 20 words in two sessions and the other consists of second-uttered ones in two sessions. The former set was used for training and the latter set for test. In the following experiments, the covariance matrix $\mathbf{\Sigma}_k$ of each Gaussian distribution $g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k)$ is assumed to be diagonal. For text-independent SV experiments, word utterances of each speaker are connected and used in an endless way. The number of tests per speaker is 20 for utterances of the same speaker and 20 for those of impostors, i.e., 2000 in total for both cases. In each test, starting point of input is

Table 1 Equal error rates (EERs) in SV experiments using GMMs with several numbers of mixtures.

	the number of mixtures		
	8	16	32
EER (%)	4.2	3.2	2.9

Table 2 Equal error rates (%) in SV experiments using artificial cohorts generated with different values of α and β.

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.5$
$\beta = 1$	10.8	10.1	11.8	–
$\beta = 2$	6.9	5.8	5.8	8.3
$\beta = 3$	7.3	6.1	5.8	6.5
$\beta = 4$	7.3	6.3	6.1	6.1

randomly selected and impostors are also randomly selected from 99 speakers excluding the relevant true speaker. SV performance is measured by equal error rate (EER).

In order to determine an appropriate number of mixtures for GMM, preliminary SV experiments were performed using universal background model. The universal background model was here estimated by using training data from all 100 speakers. Results using 50 frames are shown in Table 1. Considering that the difference of EERs for 16 and 32 mixtures is small, 16-mixture GMMs are hereafter used.

In order to determine appropriate values of α in (12) and β in (13), preliminary SV experiments were performed using artificial cohorts generated with different values of α and β. Results using 50 cohorts are shown in Table 2. It is shown that β should be greater than 1, i.e., variances for cohorts should be increased from those for the claimed speaker.[†] Considering the results in Table 2, $\alpha = 0.2$ and $\beta = 2$ are hereafter used.

The proposed SV method using artificial cohorts is evaluated by comparing it with several conventional methods: methods using the universal background model, real cohorts from the data-set, and a background model in [5], which we call pseudo background model. Additionally, SV experiments using a method without background model were also carried out where only the log-likelihood for a claimed speaker $\log p_s(\mathbf{Y})$ is used instead of the log-likelihood ratio $S(\mathbf{Y})$ in (1). In the use of cohorts, we use T-norm as well as the cohort normalization. According to our preliminary experiments shown in Table 3, the number of cohorts used are 50 for both real and artificial cohorts. Regarding the method using pseudo background model [5], 32-mixture GMM and 16-mixture GMM are recommended for a claimed speaker model and its pseudo background model, respectively. However we here use 32-mixture GMM for the claimed speaker model and 8-mixture GMM for the pseudo background model, since according to our preliminary experiments, this combination was best among combinations of 32 and 16, 32 and 8, and 16 and 8 mixtures for the claimed speaker and its pseudo background model. Experimental results are shown in Table 4. It is seen that EERs by the proposed method using artificial cohorts are about 60% less than those by the method using the pseudo background model and are about 20% less than those by the method without background model. SV performance by the proposed method using artificial cohorts is naturally worse than that using real cohorts. However, the difference in EERs between using artificial and using real cohorts can be decreased by increasing the number of frames used.

[†]We confirmed that $p_o(\mathbf{Y})$ in (7) for impostors is very small with $\beta = 1$ compared with other values of β. It is considered that $p_o(\mathbf{Y})$ generated with $\beta = 1$ cannot cover a wide range of feature vectors from impostors.

Table 3 Equal error rates (%) in SV experiments using different numbers of real and artificial cohorts.

	the number of cohorts			
	10	30	50	70
real cohorts	3.1	2.4	2.3	2.2
artificial cohorts	6.5	5.9	5.8	5.8

Table 4 Equal error rates (%) by several SV methods with different numbers of frames.

	the number of frames				
method	50	100	150	200	400
universal background	3.2	2.0	1.5	1.5	1.3
cohort (real)	2.3	1.7	1.4	1.3	1.4
T-norm (real)	2.5	1.9	1.5	1.5	1.5
cohort (artificial)	5.8	4.1	3.5	2.9	2.5
T-norm (artificial)	6.6	4.2	3.2	2.6	2.1
pseudo background	15.6	11.1	8.7	8.1	5.1
without background	6.8	5.1	4.7	4.4	3.2

5. Conclusions

In this paper, we have proposed a GMM-based text-independent SV method using only utterances of enrolled speakers, which means that it does not need speaker databases including utterances of many other speakers. Artificial cohorts are used instead of those from speaker databases, and GMMs for artificial cohorts are generated by changing model parameters of the GMM for a claimed speaker. In comparison with a conventional method and a method without background model which also do not need speaker databases, EERs by the proposed method are about 60% less than those by the conventional method and are about 20% less than those by the method without background model. SV performance by the proposed method using artificial cohorts is naturally worse than that using real cohorts. However, the difference in EERs between using artificial and using real cohorts can be decreased by increasing the number of frames used.

References

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process., vol.10, no.1-3, pp.19–41, 2000.

[2] A.E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F.K. Soong, "The use of cohort normalized scores for speaker verification," Proc. ICSLP-92, pp.599–602, 1992.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digit. Signal Process., vol.10, no.1-3, pp.42–54, 2000.

[4] O. Siohan, C.-H. Lee, A.C. Surendran, and Q. Li, "Background model design for flexible and portable speaker verification systems," Proc. ICASSP-99, pp.825–828, 1999.
[5] D. Tran and D. Sharma, "New background speaker models and experiments on the ANDOSL speech corpus," Lect. Notes Comput. Sci., vol.3214, pp.498–503, 2004.
[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc., vol.39, no.1, pp.1–38, 1977.
[7] G.J. McLachlan and K.E. Basford, Mixture models — Inference and applications to clustering, Marcel Dekker, 1988.
[8] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process., vol.3, no.1, pp.72–83, Jan. 1995.
[9] Y. Lind, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol.28, no.1, pp.84–95, 1980.
[10] H. Noda, K. Harada, and E. Kawaguchi, "A context-dependent sequential decision for speaker verification," IEICE Trans. Inf. & Syst., vol.E82-D, no.10, pp.1433–1436, Oct. 1999.