6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TEXT-INDEPENDENT SPEAKER VERIFICATION USING VIRTUAL SPEAKER BASED COHORT NORMALIZATION

Toshihiro Isobe and Jun-ichi Takahashi

Laboratory for Information Technology, NTT DATA CORPORATION
Kayaba-ch,o Tower 7F, 1-21-2, Shinkawa, Ch,u,o-ku,
Tokyo 104-0033 Japan

ABSTRACT

In this paper, we propose a new score normalization method for text-independent speaker verification using GMM (Gaussian Mixture Model). In the proposed method, cohort model is designed as virtual speaker model based on the similarity of local acoustic information between the reference speaker and other customers. The similarity is determined using statistical distance between model components such as the Gaussian distributions. Therefore, synthesized cohort model is statistically close to the reference speaker model, and can provide an effective normalizing score for various observed measurements. The experimental results using telephone speech of 60 speakers showed that the proposed method is superior to the typical methods with cohort speaker model or pooled model. Equal Error Rate (EER) when using common posteriori-defined threshold value for every speakers was drastically reduced from 3.82 % (for the conventional normalization with cohort speaker model) or 10.3 % (for normalization with pooled model) to 2.50 % (for the proposed method) when cohort size is equal to three.

1. INTRODUCTION

In speaker verification, it has been well known that score normalization using likelihood ratio of the reference speaker model and speaker background model or cohort models is very effective for improving the performance [1-4]. In the typical methods, speaker background model or cohort models are determined by choosing the closest speaker model to the reference model among the other speaker models or by combining several speaker models closer to the reference model. But, the normalizing score provided by speaker background model or cohort models is not enough to make the likelihood ratio to be stable for any texts, because their models are selected on the basis of likelihood score of the utterance. To solve this problem, we have proposed a new cohort normalization to achieve stable normalizing score by an effective method of statistical cohort speaker selection [5]. In the method, cohort models are synthesized based on the similarity of components such as phonemes, states, and the Gaussian distributions of HMMs of several speakers statistically closer to the reference speaker model. The basic idea of the method is that acoustic similarity between speakers is different for constituent segments of the utterance. The method has robustness for text variation because cohort selection is based on the statistical similarity between speakers and virtual speaker models. The proposed method is also recognized as an advanced method for text-independent speaker verification aiming at high-level security.

In this paper, we will describe the application of our virtual speaker based method for text-independent speaker verification. In the following section, basic concept is described. Section 3 describes a formulation of the proposed method. Some experimental results are also described in section 4 and 5 to make the effectiveness clear.

2. BASIC CONCEPT

Figure 1 shows conceptual illustration of cohort model synthesis method we proposed. In this figure, in order to understand the concept easily, speaker is simply represented by a GMM consisting of three mixture components.

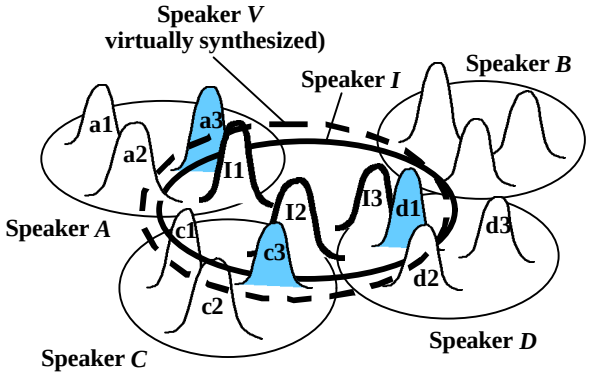


Figure 1. Concept of virtual cohort speaker model synthesis

The illustration shows the situation that four speaker models (*A*, *B*, *C*, and *D*) are closer to the reference speaker model *I*. In the conventional cohort model construction by speaker-based selection, some or all the closer models are chosen as members of cohort set. In this case, cohort set includes all distributions of speaker models A, B, C, and D. On the other hand, in the proposed method, speaker model *V* is virtually synthesized as a cohort model using some of the closer models' distributions. In this example, distribution a3 of speaker *A* is selected for

distribution I1 of reference speaker *I*. Distribution c3 of speaker *C* and distribution d1 of speaker *D* are also selected for distributions I2 and I3 of speaker *I*, respectively. These selections are proceeded based on the inter-distribution distance, which means the similarity of local acoustic features in speaker models. As shown in Fig. 1, virtually synthesized cohort model *V* is statistically closer to the reference models than cohort set or cohort model obtained by the conventional speaker-based selection. Therefore, by using our method, it is expected that the verification score represented by likelihood ratio can be less variable and more stable than the conventional one to make verification performance higher.

3. APPLICATION TO TEXT-INDEPENDENT SPEAKER VERIFICATION

In terms of log likelihood, the normalized verification score is represented as the difference of log likelihood as follows:

$$\log L(I \mid \mathbf{o}) = \log p\left(\mathbf{o} \mid f\hat{\mathbf{E}}^{(=I)}\right) - \log p\left(\mathbf{o} \mid f\hat{\mathbf{E}}^{(\neq I)}\right) \qquad (1)$$

In the typical cohort normalization methods, cohort models are organized based on speaker-based selection. Therefore, the log likelihood shown in the second term of equation (1) is represented as follows:

$$\log p\left(\mathbf{o} \mid f\hat{\mathbf{E}}^{(\neq I)}\right) = \log\left\{\frac{1}{K}\sum_{k=1}^{K} p\left(\mathbf{o} \mid f\hat{\mathbf{E}}^{(c_k(I))}\right)\right\} \qquad (2)$$

where K represents cohort size, and $c_k(I)$ is the k-th cohort speaker for the reference speaker I. A set of cohort speakers $c_k(I)$ $(k=1,2,\ldots,K)$ are selected from the registered speakers except the speaker I. The k-th selected speaker denotes the k-th closest speaker to the reference speaker I.

In the proposed virtual speaker based cohort normalization, log likelihood shown in the second term of equation (1) is represented as follows:

$$\log p\left(\mathbf{o} \mid f\hat{\mathbf{E}}^{(\neq I)}\right) = \log\left\{\frac{1}{K}\sum_{k=1}^{K} p\left(\mathbf{o} \mid f\hat{\mathbf{E}}^{(c'_k(I))}\right)\right\} \qquad (3)$$

where $c'_k(I)$ represents the k-th virtual cohort speaker.

In the text-prompt type of speaker verification using HMM as speaker model, virtual cohort speaker model is synthesized based on the local acoustic feature such as phone, state, and the Gaussian distributions of HMMs. At the phone-level selection, constituent phone of virtual speaker model is selected among some phone HMMs describing the same phoneme. For state-level, constituent state is selected from the states, which has the same number in the same phone HMMs of other registered speakers using contextual information including the text given by the system.

On the other hand, in the text-independent speaker verification using GMM, contextual information cannot be used. Constituent unit of virtual cohort speaker model is selected among all of the components of other speakers' GMMs using only acoustic information.

When speaker model is GMM described by one state with M mixture, speaker model $f\hat{\mathbf{E}}^{(I)}$ is represented as follows:

$$f\hat{\mathbf{E}}^{(I)} = \left\{a_1^{(I)}, a_{1,FIN}^{(I)}, w_m^{(I)}, \mathbf{N}_m^{(I)}\right\}_{m=1,2,\Lambda\ M} \qquad (4)$$

where $a_1^{(I)}$ is the probability of the self-loop state transition, $a_{1,FIN}^{(I)}$ is that of transition to the next model, $w_m^{(I)}$ is the weighting parameter of the m-th Gaussian distribution, and $\mathbf{N}_m^{(I)}$ denotes the m-th Gaussian distribution.

In our prposed method, GMM of the k-th virtual cohort speaker $f\hat{\mathbf{E}}^{(c'_k(I))}$ is represented as follows:

$$f\hat{\mathbf{E}}^{(c'_k(I))} = \left\{a_1^{(c'_k(I))}, a_{1,FIN}^{(c'_k(I))}, w_m^{(c'_k(I))}, \mathbf{N}_n^{(c_k(I,m))}\right\}_{m=1,2,\Lambda\ M} \qquad (5)$$

$$a_1^{(c'_k(I))} = \frac{\sum_m a_1^{(c_k(I,m))}}{\sum_m a_1^{(c_k(I,m))} + \sum_m a_{1,FIN}^{(c_k(I,m))}} \qquad (6)$$

$$a_{1,FIN}^{(c'_k(I))} = \frac{\sum_m a_{1,FIN}^{(c_k(I,m))}}{\sum_m a_1^{(c_k(I,m))} + \sum_m a_{1,FIN}^{(c_k(I,m))}} \qquad (7)$$

$$w_m^{(c'_k(I))} = w_m^{(I)} \qquad (8)$$

where $c_k(I,m)$ is the k-th cohort speaker. In the k-th cohort speaker model, the n-th Gaussian distribution is the k-th closest distribution to the m-th Gaussian distribution of GMM of speaker I. The probabilities of the state transition are re-normalized using equation (6) and (7). The weighting parameter of virtual cohort speaker is same as that of the reference speaker I.

The similarity between the Gaussian distributions is defined by the Battacharyya distance [6].

4. EXPERIMENTS

4.1 Experimental Setup

The experiments were conducted to compare the proposed method with the conventional methods such as cohort speaker model method and speaker background model method based on pooled model. Speech data were phonetically balanced and they were collected from 60 ordinary peoples. The speech data were divided into three kinds of data sets (set R, E, and X). The number of speakers was 28 (14 males and 14 females) for set R

and E, and was 32 (16 male and 16 female) for set X. For set R and E, same speakers were used, but different speakers were used for set X as an open test data. For set R, 10 utterances per speaker were collected, while 50 utterances per speaker were collected for set E and X. The data was telephone speech quality. All the speech data were recorded on the digital audio tape with headset microphone in the soundproof room at the same session. They were recorded again through practical telephone networks via mouth simulator, electret telephone handset, and DSP-based speech processing card provided by Dialogic Co.. Collected telephone speech were digitized at an 8 kHz sampling rate using an 8-bit ì-law codec, and then converted to linear PCM samples. The digitized speech signal was pre-emphasized using the filter $H(z)=1-0.95z^{-1}$, and converted to 10-th order auto-correlation coefficients in the conditions of Hamming window length: 25 msec and window shift length: 10 msec. We used 12-th order LPC cepstral coefficients, 12-th order delta cepstral coefficients, and delta log power as feature vector. Speaker models were trained by Maximum Likelihood (ML) estimation, in which each GMM has 64 mixture components. The model structure of the virtual cohort speaker models and conventional cohort speaker model was same as that of reference speaker model. Pooled model was trained using the speech data consisting of speech data of the closest speaker set to the reference speaker, in which mixture size of pooled model was 256.

4.2 Verification Experiments

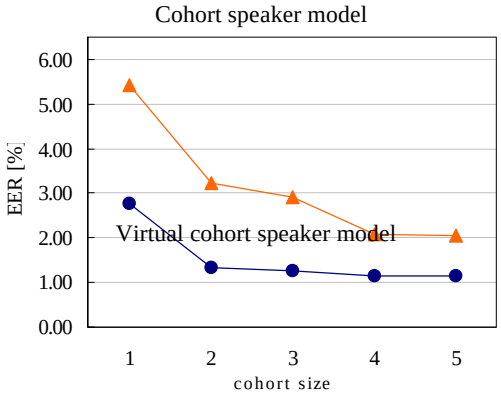
Two kinds of verification experiments were carried out: closed test and open test. In the closed test, the experiment was conducted using data sets R and E. Set R was used for training models of individual reference speakers and set E was used as speech data for verification trials. For each reference speaker, verification test was done in the assumption that other 27 speakers were recognized as imposters. Cohort models and pooled models for individual reference speakers were constructed using the other 27 speaker models and speech, respectively. In open test, individual reference models and cohort models were also obtained by the same manner as the closed test. Speakers of the data set X were used as imposters. Fifty utterances per speaker of set X were also used for verification trials.

5. RESULTS

5.1 Closed Test Results

The results of closed test are shown in Figure 2. In the figure, cohort size represents the number of cohort speakers or virtual cohort speakers. The verification performance of each method is evaluated in EER (Equal Error Rate) in the condition of posteriori-defined threshold value common to all the speakers. In Figure 2, EER decreases as cohort size increases. The performance of the conventional cohort speaker method tends to saturate when cohort size is about four to five, while that of our proposed method tends to saturate when cohort size is about two to three.

For every cohort size, the proposed method is found to be superior to the conventional one in verification performance. Furthermore, the saturation point in cohort size of the virtual cohort speaker method is smaller than that of the conventional method. It is considered that more statistically precise cohort



model can be synthesized using smaller number of distributions because of acoustically fine-level distribution selection. Compared typical method of speaker-based selection with distribution-based selection, error reduction rate of EER is 57.0 % (EER reduces from 2.91 % to 1.25 %.), when cohort size is three.

Figure 2. Closed test results

5.2 Open Test Results

In this test, speakers from data set X were used as imposters in verification test. Reference models were trained using data set R. This test seems to be practical situation in the real-world use of speaker verification system because unexpected imposters appear. Experimental results are shown in Figure 3. The similar tendency for verification performance can be found in terms of performance vs. cohort size. Cohort size is about four to five when the performance of the proposed method saturates. The high performance can be also achieved by the virtual cohort speaker method as well as the closed test. In the comparison between virtual cohort speaker method and the conventional cohort speaker method, high EER error reduction rate of 34.6 % (3.82 % to 2.50 %) can be achieved when cohort size is three.

Cohort speaker model

Virtual cohort speaker model

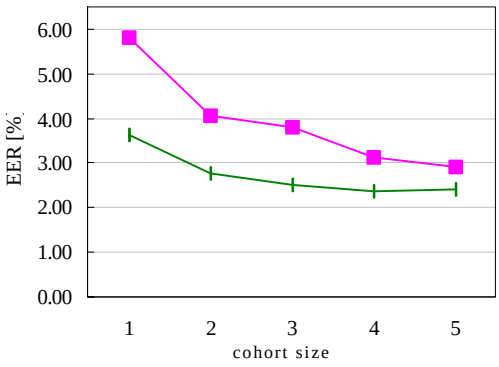


Figure 3. Open test results

5.3 Comparison Between Virtual Cohort Speaker Model and Pooled Model

In both experiments (closed test and open test), the performance of virtual cohort speaker method was compared with the method using pooled model. The result of the closed test is shown in Table 1, and that of open test is in Table 2. In the evaluation of the pooled model, cohort size is the number of speakers whose speech data are used to estimate speaker background model. In Tables 1 and 2, PM, CS, and VCS represent pooled model method, conventional cohort speaker model method, and virtual cohort speaker model method, respectively.

The performance of pooled model method is not so good as that of two kinds of cohort speaker methods. In the pooled model method, the difference of EERs is not found for two kinds of cohort sizes (size 3, and 5) in both experiments. The performance of virtual cohort speaker model method is the most effective in three kinds of score normalization methods.

Table 1. Comparison of three methods in closed test (EER)

Cohort size	PM	CS	VCS
3	9.32 %	2.91 %	1.25%
5	10.7 %	2.05 %	1.15 %

Table 2. Comparison of three methods in open test (EER)

Cohort size	PM	CS	VCS
3	10.3 %	3.82 %	2.50 %
5	10.3 %	2.89 %	2.40 %

6. DISCUSSION

At first, we discuss about comparison between virtual cohort speaker model method and conventional cohort speaker model method. From the results of the closed test and open test, our proposed method was experimentally proven to be effective for constructing cohort models. The reason is that statistical matching between cohort models and the reference models can be carried out efficiently by finely evaluating local acoustic similarity based on the statistical difference between the Gaussian distributions. It can be said that our proposed method is effective for text-independent speaker verification, which has large contextual variation.

Secondly, we conclude the relation between EER and cohort size. From Figures 2 and 3, it is found that EER decreases and saturates as cohort size increase. The reason is that the closer components to the reference model are dominant for normalizing score in cohort speaker models or virtual cohort speaker models. Even if the number of components which are statistically far from the reference model increases, they do not have any influence on score normalization. EER of our proposed method saturates early than that of conventional method. It is because the proposed method finely selects the components of virtual cohort speaker models based on the local acoustic information. Therefore, the method cans combine optimal unit to cohort models at each cohort size.

Finally, we discuss about the comparison between cohort speaker model and pooled model. In both experiments (closed test and open test), EER of the pooled model method increased as compared with that of the methods based on cohort speaker model. In text-independent speaker verification, the pooled model represents acoustic feature of all phonemes of several speakers by one model, so it is considered that the statistic distributions of the pooled model become broad and the score normalizing ability is reduced. In order to analyze this assumption, the experimental evaluation might be done by phoneme-based pooled model as our future work.

7. SUMMARY

In this paper, we proposed a new cohort normalization for text-independent speaker verification using GMM. The proposed method uses virtual cohort speaker models synthesized on the basis of local acoustic features represented by the Gaussian distributions of GMMs. Cohort models obtained by the method can provide an effective normalizing score when verification is carried out using various observation sequences. The reason is that the synthesized models are statistically close to the reference models. Some experimental results showed that our proposed method is effective for score normalization and has much robustness for contextual variation, because grain of constituent unit for synthesizing cohort models is so fine to control the normalizing score. The experimental results showed that the proposed method is superior to the other methods. Equal Error Rate (EER) by posteriori-defined threshold was drastically reduced from 3.82 % (obtained by the conventional normalization with cohort speaker model) or 10.3 % (obtained by normalization with pooled model) to 2.5 % (obtained by the proposed method) when cohort size is equal to three. High EER error reduction rates of 34 % and 42 % were achieved. The effectiveness of our proposed method was experimentally

proven for contextual variation of input utterances in text-independent speaker verification.

8. REFERENCES

[1] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Processing*, vol. 1, pp.89-106, 1991.

[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*, vol. 2, pp.599-602, 1992.

[3] C-S. Liu, H-C. Wang, and C-H. Lee, "Speaker Verification Using Normalized Log-Likelihood Score," *IEEE Transactions on Speech and Audio Processing*, vol. 4, No. 1, pp.56-60, 1996.

[4] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition," *Proc. ICASSP 93*, vol. 2, pp.391-394, 1993.

[5] T. Isobe and J. Takahashi, "A New Cohort Normalization Using Local Acoustic Information for Speaker Verification," *Proc. ICASSP 99*, vol. 2 pp.150-153, 1999.

[6] K. Fukunaga, "Introduction on Statistical Pattern Recognition (Second Edition), " *Academic Press, Inc., San Diego*, 1990.