A NEW COHORT NORMALIZATION USING LOCAL ACOUSTIC INFORMATION FOR SPEAKER VERIFICATION

Toshihiro Isobe and Jun-ichi Takahashi

Laboratory for Information Technology, NTT DATA CORPORATION
66-2, Horikawa-cho, Saiwai-ku, Kawasaki-shi,
Kanagawa 210-0913 Japan

ABSTRACT

This paper describes a new cohort normalization method for HMM based speaker verification. In the proposed method, cohort models are synthesized based on the similarity of local acoustic features between speakers. The similarity can be determined using acoustic information lying in model components such as phonemes, states, and the Gaussian distributions of HMMs. With the method, the synthesized models can provide an effective normalizing score for various observed measurements because the difference between the individual reference model and the synthesized cohort models is statistically reduced through fine evaluation of acoustic similarity in model structure level. In the experiments using telephone speech of 100 speakers, it was found that high verification performance can be achieved by the proposed method: the Equal Error Rate (EER) was drastically reduced from 1.20 % (obtained by the conventional speaker-selection based cohort normalization) to 0.30 % (obtained by the proposed method on distribution-based selection) in closed test. Furthermore, EER was also reduced from 1.40 % to 0.70% in open test (reference speaker: 25, impostor: 75), when the other speakers than the reference speaker were used as impostors.

1. INTRODUCTION

In most speaker verification methods, it has been known that score normalization using the likelihood ratio of the reference speaker model and speaker background model or cohort model is very effective for improving the performance. Higgins *et al.* used a discriminate counter to verify the speakers, in which they used a maximum score of all speaker models (the reference model is not included) as normalizing score[1]. Rosenberg *et al.* used a set of models consisting of several cohort speakers selected for individual reference speakers. In the study, cohort set was constructed by random selection of speakers, selection based on similarity between speakers, and so on[2]. Liu *et al.* proposed the use of cohort models given by pooled training, in which speech data of the speakers whose models are similar to the reference model are used[3]. Furthermore, Matsui *et al.* reported the effectiveness of the normalization based on a posteriori probability, in which the reference speaker is included in training of cohort models[4]. In terms of log likelihood, the normalized verification score is represented as the difference of log likelihoods, as follows:

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

where **o** represents the observed sequence of feature vectors, and $p\left(\mathbf{o} \mid \lambda^{(s=I)}\right)$ is the likelihood of the observed sequence with respect to the reference speaker I, and $p\left(\mathbf{o} \mid \lambda^{(s\neq I)}\right)$ is the likelihood of the sequence for other speakers than I. According to the equation (1), the key point is how to construct cohort models $\lambda^{(s\neq I)}$ which provide an effective normalizing score robustly against the likelihood variation for various observation sequences. In the above-mentioned normalization methods, various techniques were devised to construct cohort models. But there is serious problem in these methods. In the methods, cohort models are determined by choosing the closest speaker model to the reference model among the other speaker models or combining some speaker models closer to the reference model. Constituent unit of cohort set is obliged to be "speaker model", so the likelihood variation of cohort models is difficult to control finely. Therefore, it is considered that the likelihood score ratio is not stable.

In this paper, to solve the problem, a new method of constructing cohort set and the way of synthesizing cohort models are proposed. The feature of the method is that cohort models are virtually synthesized focused on the acoustic similarity between models in fine-structure level. In the following section, basic concept is described. Section 3 describes a formulation of the proposed method. Some experiments and their results also described in section 4 and 5.

2. BASIC CONCEPT

Figure 1 shows conceptual illustration of cohort model construction method we proposed. In this figure, in order to understand the concept easily, speaker is simply represented by a model consisting of three Gaussian mixture distributions. 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. On the other hand, in distribution-based selection, which is one of the proposed methods, speaker model *V* is virtually constructed as cohort model using some of the closer models' distributions. In the 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 determined by distance between distributions, which mean the similarity of local acoustic features. 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. This means that the verification score represented by the likelihood ratio shown in equation (1) becomes less variable and more stable by the use of our method than that of conventional one. Therefore, it is expected that our method can provide effective cohort models for normalizing score.

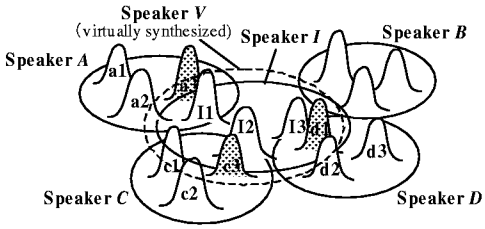

Figure 1. Concept of cohort model construction

3. COHORT NORMALIZATION TECHNEQUES USING LOCAL ACOUSTIC INFORMATION

3.1 Selection Method

(a) Speaker-based Selection

Typical cohort normalization methods are based on speaker-based selection in organizing cohort set. Log likelihood shown in the second term of equation (1) is represented as follows:

$$\log p\left(\mathbf{o} \mid \boldsymbol{\lambda}^{(s \neq I)}\right) = \log\left\{\frac{1}{K}\sum_{k=1}^{K} p\left(\mathbf{o} \mid \boldsymbol{\lambda}^{(c_k(I))}\right)\right\} \quad (2)$$

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 reference speakers except the speaker I. The k-th selected speaker is the k-th closest speaker to the reference speaker I.

(b) Phoneme-based Selection

When organizing cohort models by phoneme-based selection, the second term of equation (1) is represented as follows:

$$\log p\left(\mathbf{o} \mid \boldsymbol{\lambda}^{(s \neq I)}\right) = \log\left\{\frac{1}{K}\sum_{k=1}^{K} p\left(\mathbf{o} \mid \boldsymbol{\lambda}^{(c'_k(I))}\right)\right\} \quad (3)$$

where $c'_k(I)$ represents the k-th cohort speaker which model is virtually synthesized based on the following way of phoneme selection. The synthesized model consists of phonemes extracted from different speaker models. Each phoneme model is closer to the corresponding phoneme model of the reference speaker I. In general expression, the selection can be represented in the following. Given a set of models of speaker I as

$$\boldsymbol{\lambda}^{(I)} = \left\{\lambda_1^{(I)}, \lambda_2^{(I)}, \cdots, \lambda_p^{(I)}, \cdots, \lambda_P^{(I)}\right\} \quad (4)$$

where P is the number of phoneme models and $\lambda_p^{(I)}$ is a model represented by HMM for phoneme p of speaker I, a set of phoneme models of virtually synthesized k-th cohort speaker is described as

$$\boldsymbol{\lambda}^{(c'_k(I))} = \left\{\lambda_1^{(c_k(I,1))}, \lambda_2^{(c_k(I,2))}, \cdots, \lambda_p^{(c_k(I,p))}, \cdots, \lambda_P^{(c_k(I,P))}\right\} \quad (5)$$

where $\lambda_p^{(c_k(I,p))}$ is a model obtained from phoneme p of the k-th cohort speaker, which corresponds to the model for phoneme p of the reference speaker I.

(c) State-based Selection

In the case of cohort models by state-based selection, the likelihood is represented as equation (2). Topology of phoneme HMMs is assumed to be left-to-right, in which each HMM consists of S states with a mixture of M Gaussian distributions per state in the following:

$$\lambda_p = \{a_{p,s,s}, a_{p,s,s+1}, w_{p,s,m}, N_{p,s,m}\}_{s=1,2,\cdots S;\, m=1,2,\cdots M;} \quad (6)$$

where p represents phoneme, $a_{p,s,t}$ is the probability of state transition from state s to t, $w_{p,s,m}$ is the weighting parameter of the m-th Gaussian distribution of state s, and $N_{p,s,m}$ denotes the m-th Gaussian distribution of state s. Therefore, a set of models of virtually synthesized k-th cohort speaker is represented as follows:

$$\boldsymbol{\lambda}^{(c'_k(I))} = \{\boldsymbol{a}_{p,s,s}^{(c_k(I,p,s))}, \boldsymbol{a}_{p,s,s+1}^{(c_k(I,p,s))}, \boldsymbol{w}_{p,s,m}^{(c_k(I,p,s))}, \boldsymbol{N}_{p,s,m}^{(c_k(I,p,s))}\}_{p=1,2,\cdots P;\, s=1,2,\cdots S;\, m=1,2,\cdots M;} \quad (7)$$

where $c_k(I,p,s)$ is the k-th cohort speaker selected when state s of phoneme model p is the k-th closest state to the corresponding state of the same phoneme model of speaker I.

(d) Distribution-based Selection

For distribution-based selection, a set of models for virtually synthesized k-th cohort speaker is defined as follows:

$$\boldsymbol{\lambda}^{(c'_k(I))} = \{\boldsymbol{a}_{p,s,s}^{(c'_k(I))}, \boldsymbol{a}_{p,s,s+1}^{(c'_k(I))}, \boldsymbol{w}_{p,s,m}^{(c'_k(I))}, \boldsymbol{N}_{p,s,n}^{(c_k(I,p,s,m))}\}_{p=1,2,\cdots P;\, s=1,2,\cdots S;\, m=1,2,\cdots M;} \quad (8)$$

$$\boldsymbol{a}_{p,s,s}^{(c'_k(I))} = \frac{\sum_m \boldsymbol{a}_{p,s,s}^{(c_k(I,p,s,m))}}{\sum_{j=0,1}\sum_m \boldsymbol{a}_{p,s,s+j}^{(c_k(I,p,s,m))}} \quad (9)$$

$$\boldsymbol{w}_{p,s,m}^{(c'_k(I))} = \frac{\boldsymbol{w}_{p,s,n}^{(c_k(I,p,s,m))}}{\sum_m \boldsymbol{w}_{p,s,n}^{(c_k(I,p,s,m))}} \quad (10)$$

where $c_k(I,p,s,m)$ is the k-th cohort speaker. In the k-th cohort speaker model, the n-th Gaussian distribution at state s for phoneme model p is the k-th closest distribution to the m-th Gaussian distribution at the same state of phoneme model p for speaker I. The probabilities for self-loop state transition and weighting parameter are renormalized using equations (9) and (10) according to the constraints given by mathematical HMM formulation, because the constituent Gaussian distribution has selected from different speaker models.

3.2 Similarity Measure in Each Selection

When cohort models are constructed, constituent components of the models are chosen from all the reference models based on the similarity between speaker models, phoneme models, states, and distributions. In the proposed methods, the similarity between components is defined on the basis of the Battacharyya distance

[5] of the Gaussian distributions of HMMs. Let us consider the similarity between speaker model X and speaker model Y, in which X and Y consists of the same phoneme sequences and those are for different speakers. In the Gaussian distribution-based selection, distance between distributions at the same state number in models X and Y is evaluating as the inter-distribution similarity. When using state-based selection, representative distributions are estimated using constituent mixture distributions of the state, and then the distance between them is evaluated as the inter-state similarity at the same state number in models X and Y. In the phoneme-based selection, the averaged value of whole inter-state distances is used as inter-phoneme similarity. Inter-speaker similarity is also used as the averaged distance of entire inter-phoneme distances obtained from constituent phoneme models.

4. EXPERIMENTS

4.1 Experimental Setup

Four kinds of data sets (sets A1, A2, A3, and B) were used for the experiment. Vocabulary items of those data sets were four-connected digits. The number of speakers was 25 (12 males and 13 females) for sets A1, A2, and A3, and was 75 (38 males and 37 females) for set B. In the data sets, 70 utterances of four-connected digits per speaker were collected. The data sets A1, A2, and A3 consist of speech data uttered by the same speakers but each of them were collected every three months. The data were telephone speech quality. The data sets A1 and B were recorded at the same session. All the speech data were recorded on the digital audio tape with headset microphone in the soundproof room. 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 μ-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 as context-independent phoneme HMMs, in which each HMM was 3-state let-to-right model with 3 mixture components per state.

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 A1, A2, A3, respectively. Thirty speech data of 70 utterances in data set A1 were used for training models of individual reference speakers. Forty utterances of each data sets A1, A2, and A3 were used as speech data for verification trials. For each reference speaker, verification test was done in the assumption that other 24 speakers were recognized as imposters. For data set A2 and A3, reference models obtained using data set A1 were also used for verification test. Cohort models for individual reference speakers were constructed by the proposed methods using the other 24 speaker models. On the other hand, in open test, individual reference models and cohort models were obtained by the same manner as the closed test. Speakers of the data set B were used as imposters against the data set A1, A2, and A3. Forty utterances per speaker of data set B were also used for verification trials.

5. RESULTS

5.1 Closed Test Results

The results of closed test are shown in Figs. 2-(a) to 2-(c). The verification performance of the proposed methods is evaluated in EER (Equal Error Rate) in the condition of posteriori-defined threshold value common to all the speakers. For any data set A1, A2, and A3, EER decreases as cohort size increases for every method. Performance tends to saturate when cohort size is about five to six. In the viewpoint of difference between synthesized methods, distribution-based selection was most superior to the others. Performance tends to decrease in the order of state-based selection, phoneme-based selection, and speaker-based selection. This shows the tendency that performance increases when the grain of constituent unit become so fine in cohort model construction. Compared typical method of speaker-based selection with distribution-based selection, error reduction rate of EER is 70.1 % (EER reduces from 0.87 % to 0.26 %.) for the data set A1, 37.5 % (2.88 % to 1.80 %) for A2, and 24.4 % (4.5 % to 3.4 %) for A3, when cohort size is five.

5.2 Open Test Results

In this test, speakers from data set B were used as imposters in verification test for data set A1, A2, and A3, respectively. Models trained using data set A1 were used as reference models. 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 Figs. 3-(a) to 3-(c). The similar tendency for verification performance can be found in terms of performance vs. cohort size. Cohort size is about four to five when performance saturates. The high performance can be also obtained by the distribution-based selection as well as the closed test. Performance comparison between selection methods gives the same tendency as the closed test in the tests for data set A1, A2, and A3: distribution-based, state-based, phoneme-based, and speaker-based in the order of performance. In the comparison between speaker-based selection and distribution-based selection, high EER error reduction rate of 46.2 % (1.30 % to 0.7 %) can be achieved for data set A1. The rates are 24.8 % (3.99 % to 3.0 %) for A2 and 24.0 % (6.0 % to 4.56 %) for A3, respectively. For these tests, cohort size is five.

6. DISCUSSION

From the results of closed test and open test, our proposed method was experimentally proven to be effective for constructing cohort models. The Gaussian distribution-based selection was the most effective in any other selection methods. 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 difference between the Gaussian distributions.

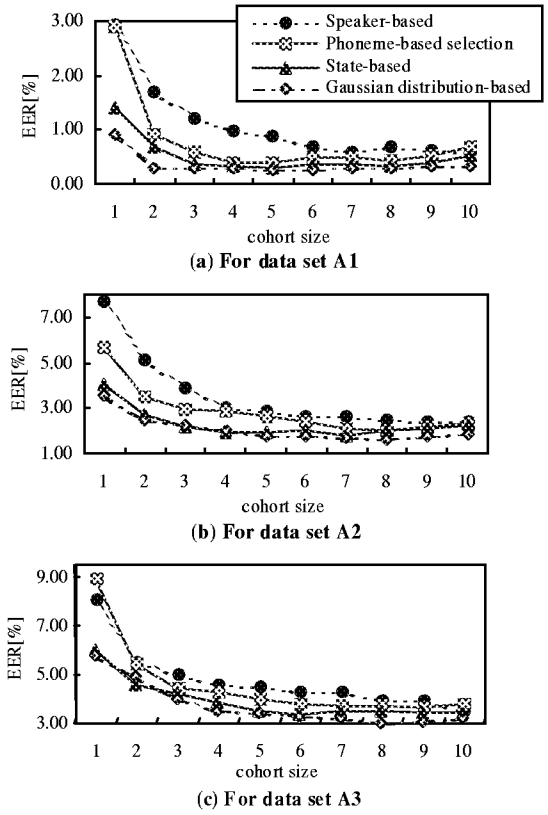

Figure 2. Closed test results

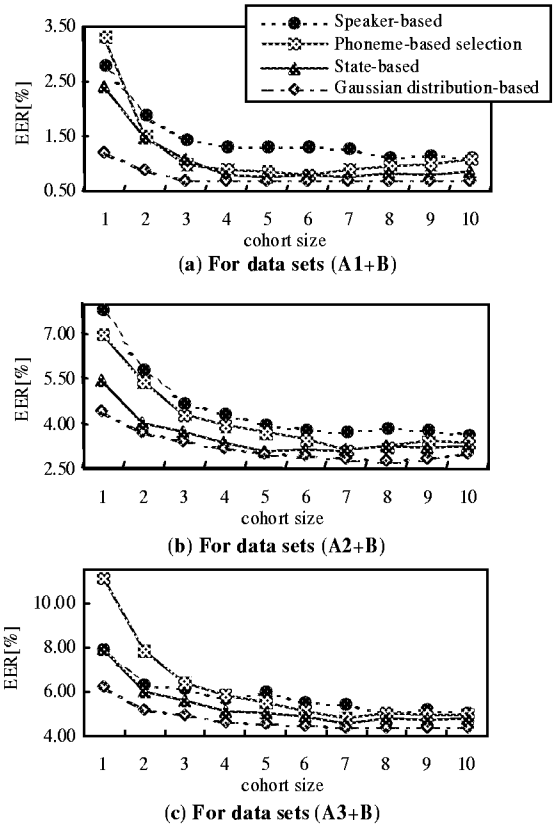

Figure 3. Open test results

We measured mean and variance for posteriori-defined threshold values of individual reference speakers. In this case, EER values are different each other reference speakers. The results are shown in Table 1. In this table, we must remark an important point that distribution-based selection can give the smallest variance of threshold values. This means that the proposed distribution-based selection can provide an effective normalizing score robustly for various measurements. Considering this result of small variance along equation (1), we can recognize that cohort models constructed by the proposed method are statistically closer to the reference models.

Table 1. Statistics of threshold values for individual speaker's EERs

	Speaker-based	Phone-based	State-based	Gaussian distribution-based
Mean	29.9	28.0	27.6	10.4
Variance	133.5	125.2	108.9	72.6

7. SUMMARY

We proposed a new cohort normalization for speaker verification. Our proposed method uses cohort models synthesized based on local acoustic features of various kind of components such as phonemes, states, and the Gaussian distributions of HMMs. Cohort models obtained by the method can provide an effective normalizing score when verification is carried out using various observation sequences. Because the synthesized models are statistically close to the reference models. Some experimental results showed that distribution-based selection is most effective, because grain of constituent unit for synthesizing cohort models is so fine to control the normalizing score variation. From some experiments, in open test of 100 speakers verification (reference speaker: 25, impostor: 75), high EER reduction rate can be achieved, compared with the conventional speaker-based selection normalization: 46.2 % for test set in which training data and trial data were recorded in the same session, 24.8 % for test data in which trial data were collected three months later, and 24.0 % for test data (recorded six months later), where cohort models are synthesized using the Gaussian distribution-based selection and cohort size is five.

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