Distributed Speaker Recognition using Speaker-Dependent VQ Codebook and Earth Mover's Distance

Shingo Kuroiwa, Yoshiyuki Umeda, Satoru Tsuge, Fuji Ren

Department of Engineering
The University of Tokushima, Tokushima, Japan
kuroiwa@is.tokushima-u.ac.jp

Abstract

In this paper, we propose a distributed speaker recognition method using a non-parametric speaker model and Earth Mover's Distance (EMD). In distributed speaker recognition, the quantized feature vectors are sent to a server. The Gaussian mixture model (GMM), the traditional method used for speaker recognition, is trained using the maximum likelihood approach. However, it is difficult to fit continuous density functions to quantized data. To overcome this problem, the proposed method represents each speaker model with a speaker-dependent VQ code histogram designed by registered feature vectors and directly calculates the distance between the histograms of speaker models and testing quantized feature vectors. To measure the distance between each speaker model and testing data, we use EMD which can calculate the distance between histograms with different bins. We conducted text-independent speaker identification experiments using the proposed method. Compared to results using the traditional GMM, the proposed method yielded relative error reductions of 85% for quantized data.

1. Introduction

In conventional mobile telephone speaker recognition systems, speech signals are encoded at the terminal side and the coded speech is transmitted to the server where the recognition system is installed. However, because there are problems of channel distortion and codec distortion, recognition performance degrades significantly.

In speech recognition, a Distributed Speech Recognition (DSR) system has been proposed to avoid these problems[1]. DSR separates the structural and computational components of recognition into two components - the front-end processing on the terminal and the speech recognition engine on the server. The European Telecommunications Standards Institute (ETSI) has published standard DSR front-end algorithms based on Mel-Cepstrum technology[2]. In the future, we expect that speaker recognition will also shift to the distributed system. One advantage of distributed speaker recognition systems is that they can use a high frequency component. This is a very important point for speaker recognition. Recently, some researchers have focused on distributed speaker recognition and have reported their results of distributed speaker recognition using ETSI standard DSR front-end[3, 4, 5, 6].

Distributed systems compress the sending data by establishing a lower bit rate for transmission. The ETSI standard DSR front-end employs a split vector quantization (SVQ) algorithm for this compression algorithm. *Fukuda et al.* have reported that the quantized data negatively affect recognition performance[3]. *Chin-Hung Sit et al.* have also reported that it is difficult to use the maximum likelihood approach (based on the EM algorithm) to train a Gaussian mixture model (GMM) whose output probability is represented with a continuous density function to fit the quantized data [5]. If unquantized feature vectors are used to train the speaker model, we can avoid this problem. However, unquantized data can not be obtained in a distributed environment.

To investigate the reason behind recognition performance degradation, we conducted the speaker recognition experiment described in section **2**. From this investigation, we concluded that it is difficult to estimate the variance of GMM using the quantized feature vectors because many variance elements are floored.

In this paper, we propose a novel non-parametric speaker recognition technique which does not require estimating statistics parameters of the speaker model. We represent a speaker model using a histogram of speaker-dependent VQ codebook. Using this speaker model, we can avoid the problem of estimating the variance. To calculate the distance between this speaker model and the recognized feature vectors, we apply a Earth Mover's Distance (EMD) algorithm. The EMD algorithm has been applied to calculate the distance between two image data represented by histograms[1] of multidimensional features[8]. Since the proposed method does not need estimation of statistics parameters, it is expected that the proposed method is robust to the quantized data. Although we have already proposed the EMD-based distributed speaker recognition that uses the ETSI speaker-independent VQ codebook in ICSLP2004[9], we show some improvement by using speaker-dependent VQ codebook and a comparison with the conventional VQ-distortion method in this paper.

Section **2** describes the distributed speaker recognition system and the influence of feature compression based on ETSI standard DSR front-end. Section **3** explains the proposed speaker identification method, and section **4** presents our experiments for speaker identification. Finally, we summarize this paper in section **5**.

2. Problems of GMM under the condition of distributed environment

In this section, we first describe the distributed speaker recognition paradigm and the influence of compression. Next, we conduct a speaker recognition experiment using GMM to investigate the influence of quantized feature vectors.

[1] In [8], EMD is defined the distance between two *signatures*. The *signatures* are histograms that have different bins, so that we use "histogram" as a term in this paper.

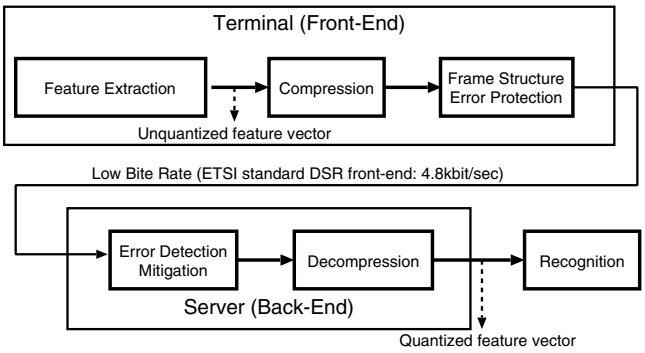

Figure 1: A block diagram of distributed speaker recognition system

2.1. Distributed speaker recognition

It is expected that the distributed speaker recognition system follows the block diagram shown in Fig. 1. This block diagram is almost the same as for the distributed speech recognition. The paradigm of distributed speech recognition has become standardized. In actually, the front-end of the distributed speech recognition has been recommended by ETSI. Although the distributed speaker recognition standard has not been recommended yet, in this paper, we use the ETSI standard DSR front-end, ES 201 108, for distributed speaker recognition front-end. In DSR, many researchers use the 8kHz standard DSR front-end. The 8kHz sampling speech data have been used in the AURORA2 database[7] which is a standard database for evaluating the DSR methods. Nevertheless, in speaker recognition, we should use higher sampling frequencies to improve recognition accuracy. Accordingly, in this paper, we try to use the 16kHz standard whose transmitting bit rate is the same as for the 8kHz standard, i.e., 4.8kbps.

The ETSI standard DSR front-end compresses the feature vectors for transmitting over the network. As a compression method, ETSI employed split vector quantization. Since the feature vectors at the server-side are quantized data, the distribution of the quantized data becomes discrete.

2.2. Speaker identification experiments using GMM

We conducted the speaker identification experiment using GMM to investigate the influence of feature compression. The training data and the conditions of acoustic analysis are explained in section **4.1**. We used the 16-mixture GMMs in this experiment.

Table 1 shows the experimental results. From this table, we can see that the performance of quantized feature vectors at 16kHz is better than for unquantized feature vectors at 8kHz. This result indicates that the distributed speaker recognition system can improve recognition performance compared with the traditional telephone speaker recognition systems. From this result, we used the 16kHz sampling speech data in the following sections.

Comparing the performances of the quantized feature vector and the unquantized feature vector in the table, we can observe that the quantization of feature vectors negatively affects the recognition performances. We consider that the reason is that the dispersion of the feature vectors by the vector quantization negatively affects the speaker model training. In fact, we

Table 1: Error rate (IER) of text-idependent speaker identification experiments at 8 and 16 kHz

	Sampling rate	
feature vector	8kHz	16kHz
Unquantized	7.0 %	3.3 %
Quantized	23.5 %	4.0 %

observe that many variance elements are floored when we investigate the speaker models. Hence, it is difficult to estimate the variance, which is a statistical parameter, using the quantized feature vector. To overcome this problem, *Chin-Hung Sit* et.al. [5] proposed to add zero-mean, random vectors to the quantized MFCCs to produce the training vectors. On the other hand, we try to use the non-parametric speaker model instead of the parametric speaker model like GMM. In the following section, we describe a proposed speaker recognition method using a non-parametric speaker model representation and EMD.

3. Non-parametric speaker recognition method using EMD

In this section, we propose a distributed speaker recognition method using a non-parametric speaker model and EMD. The proposed method uses a histogram of speaker-dependent VQ codebook as the non-parametric speaker model and calculates the EMD between speaker model and testing feature vectors. Since we use a histogram of VQ codebook, we avoided the estimation problem of statistical parameters, variance flooring. The EMD algorithm is used for directly calculating the distance between histograms that have different bins.

3.1. Earth Mover's Distance

The EMD was proposed by *Rubner et al.*[8] for an efficient image retrieval method. In this section, we describe the EMD algorithm according to their paper.

The EMD is defined as the minimum amount of work needed to transport *goods* from several *suppliers* to several *consumers*. The EMD computation has been formalized by the following linear programming problem: Let $\boldsymbol{P} = \{(\boldsymbol{p}_1, w_{p_1}), \ldots, (\boldsymbol{p}_m, w_{p_m})\}$ be the discrete distribution, such as a histogram, where $\boldsymbol{p}_i$ is the centroid of each cluster and w_{p_i} is the corresponding weight (=frequency) of the cluster; let $\boldsymbol{Q} = \{(\boldsymbol{q}_1, w_{q_1}), \ldots, (\boldsymbol{q}_n, w_{q_n})\}$ be the histogram of test feature vectors: and $\boldsymbol{D} = [d_{ij}]$ be the ground distance matrix where d_{ij} is the ground distance between centroids $\boldsymbol{p}_i$ and $\boldsymbol{q}_j$.

We want to find a flow $\boldsymbol{F} = [f_{ij}]$, with f_{ij} the flow between $\boldsymbol{p}_i$ and $\boldsymbol{q}_j$, that minimizes the overall cost

$$WORK(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{F}) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}, \qquad (1)$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad (1 \leq i \leq m, 1 \leq j \leq n), \quad (2)$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \qquad (1 \leq i \leq m), \quad (3)$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j} \qquad (1 \leq j \leq n), \quad (4)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_i}\right). \quad (5)$$

Constraint (2) allows moving *goods* from $\boldsymbol{P}$ to $\boldsymbol{Q}$ and not vice versa. Constraint (3) limits the amount of *goods* that can be sent by the cluster in $\boldsymbol{P}$ to their weights. Constraint (4) limits the amount of *goods* that can be received by the cluster in $\boldsymbol{Q}$ to their weights. Constraint (5) forces to move the maximum amount of *goods* possible. They call this amount the *total flow*. Once the transportation problem is solved, and we have found the optimal flow $\boldsymbol{F}$, the EMD is defined as the work normalized by the total flow:

$$EMD(\boldsymbol{P}, \boldsymbol{Q}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (6)$$

The normalization factor is the total weight of smaller distribution, because of constraint (5). This factor is needed when the two distributions have different total weight, in order to avoid favoring smaller distribution.

3.2. The recognition flow of the proposed method

In the previous section, we described that EMD is calculated as the least amount of work which fills the requests of *consumers* with the goods of *suppliers*.

If we define the speaker model as the *supplier* and the testing feature vectors as the *consumer*, the EMD can be applied to sparker recognition. Hence, we propose a distributed speaker recognition method using a non-parametric speaker model and EMD measurement. The proposed method represents the speaker model and testing feature vectors as a histogram. Fig. 2 illustrates the flow of the proposed method. The detail of the proposed method is described as follows:

- Speaker model generation:

  The speaker recognition system obtains each speaker's quantized feature vectors extracted using ETSI standard DSR front-end for generating a speaker model. Using these feature vectors, the system generates each speaker's VQ codebook and then makes a histogram of this VQ codebook. The histogram is used as a speaker model which is the *supplier*'s discrete distribution, $\boldsymbol{P}$, described in the previous section.

- Testing data:

  The testing feature vectors are extracted and quantized using ETSI standard DSR front-end. Using these quantized feature vectors, the system makes a histogram of VQ codebook that is employed by ETSI standard DSR front-end. This histogram is used as the *consumer*'s discrete distribution, $\boldsymbol{Q}$, described in the previous section.

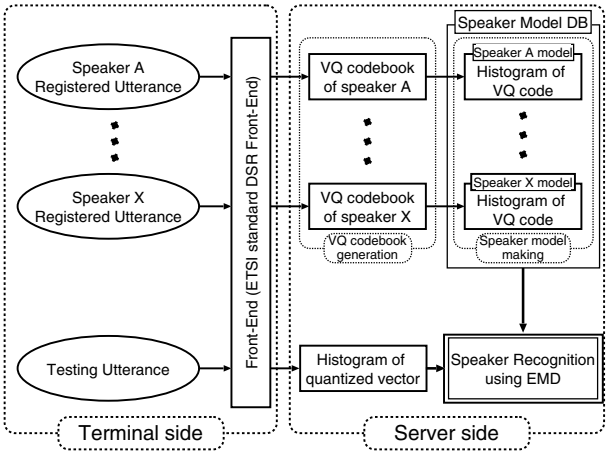

Figure 2: The block diagram of proposed method

- Identification:

  The system calculates the distance between the histogram of each speaker model and the histogram of the testing data. Then, the system selects the speaker model that has the minimum distance. This procedure requires a distance measurement between histograms which have different bins. We apply the EMD algorithm to calculate this distance. In the proposed method, the weight value, w_i, is equivalent to the occurrence frequency of a corresponding VQ centroid and the grand distance, d_{ij}, is the Euclidean distance in MFCC vector space.

4. Experiments

We conducted text-independent speaker identification experiments to evaluate the proposed method using a de facto standard Japanese speech database for speaker recognition.

4.1. Experimental conditions

From the database, we use 21 male speakers' utterances. These utterances are recorded in 7 sessions over 19 months, from Aug. '90 to Mar. '92. Each speaker spoke ten text sentences, the average length of which is five seconds.

For the registered data, i.e., the speaker model training data, we used five text sentences which were uttered in Aug. '90 by all speakers. The utterances of the remaining six sessions were used for testing. For the test set, we used five text sentences in six sessions by all speakers, in total 630 utterances (21 speakers $\times$ 5 sentences $\times$ 6 sessions). The text of these utterances is not contained in the training data [2].

These utterances, sampled at 16kHz, were segmented into overlapping frames of 25ms, producing a frame every 10ms. A Hamming window was applied to each frame. Mel-filtering was performed to extract 12-dimensional static Mel-Frequency Cepstral Coefficients (MFCC), as well as a logarithmic energy measure in the DSR front-end. The twelve dimensional delta MFCC were extracted from the static MFCC received at the server to constitute a feature vector of 25 MFCC's (12 static MFCCs extracted from the DSR front-end + 12 delta MFCC +

[2] In [9], we used 10 sentences $\times$ 6 sessions. The text of five sentences in them was contained in training data.

Table 2: Error rate (IER) of text-indepedent speaker identification experiments

Method	IER
GMM	4.0 %
VQ-distortion	1.0 %
EMD (ICSLP2004)	1.0 %
EMD-VQ	0.6 %

delta log-power). Cepstral Mean Subtraction (CMS) was applied on static MFCC vectors.

In this experiment, we set the number of centroids to 256. For comparison with the proposed method, we also conducted experiments with the speaker recognition methods based on GMM, VQ-distortion, and EMD using speaker-independent VQ codebook which was propsed in ICSLP2004. The GMM with 64 mixture was trained with the same feature vectors. The codebook of VQ-distortion method was the same as the proposed method. Each of methods indicated the best accuracy using these parameters.

4.2. Experimental results

Table 2 shows the speaker identification error rate (IER) obtained using the proposed method (EMD-VQ). For comparison, we also show the IER obtained using GMM, VQ-distortion, and EMD using speaker-independent VQ codebook (EMD) in the table.

We can see from these results that the VQ-distortion method had a lower IER than the GMM method; 1.0% for VQ-distortion and 4.0% for GMM. From this result, we conclude that the VQ-distortion method, which is a non-parametric technique, is a better method for modeling quantized data than GMM. When we investigated each speaker's GMM, we observed that many variance elements were floored. This estimation would negatively affect recognition performance. Thus, we conclude that non-parametric modeling is the proper method for distributed speaker recognition.

In addition, we can also see from this table that the proposed method, EMD-VQ, decreased the IER over the VQ-distortion method. Although these two methods, EMD-VQ and VQ-distortion, use the same VQ codebook, the EMD-VQ method shows a lower IER than the VQ-distortion method. We expect the reason for this result is the difference of distance measures. The proposed method directly calculates the distance between histograms, while the VQ-distortion method calculates the distance by totaling the VQ distortion of each frame. The proposed method can compare the distribution of speaker model with the distribution of the testing feature vectors. Therefore, the proposed method improved the IER of the VQ-distortion method. Therefore, we conclude that the EMD is a useful distance measure for speaker recognition.

On the other hand, from the comparison of the results of EMD-VQ and EMD, we may conclude that it is better to use a speaker-dependent VQ codebook rather than a speaker-independent one. Althoug, we have to conduct further investigation to clarify the cause of this fact, we conclude that EMD-VQ is an effective method for distributed speaker recognition.

5. Summary

In this paper, we have proposed a novel distributed speaker recognition method using a non-parametric speaker model and Earth Mover's Distance (EMD). To avoid the problem of estimating the variance of GMM, the proposed method represents the speaker model with the histogram of speaker-dependent VQ codebook. In the proposed method, testing data are also represented with the histogram of ETSI DSR standard codebook. To calculate the distance between the speaker model and the testing data, we have applied EMD which calculates the distance between histograms with different bins.

Experimental results on Japanese speaker identification showed that the proposed method gave a consistently better performance than the conventional methods, GMM and VQ-distortion. We confirmed improvements in the identification error rate (85% over the GMM method and 40% over the VQ-distortion method) by using the proposed method.

In future work, we will evaluate the performance of the proposed method using a larger database, and apply the proposed method to speaker verification.

6. Acknowledgments

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 15700163, Grant-in-Aid for Exploratory Research, 17656128, Grant-in-Aid for Scientific Research (B), 14280116, 17300065, 2005, and International Communications Foundation (ICF).

7. References

[1] Lilly, B. and Paliwal, K., "Effect of speech coders on speech recognition performance," *ICSLP96*, pp. 2344–2347, Oct. 1996.

[2] ETSI ES 201 108 v1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithm.

[3] Fukuda, I., Fattah, M. A., Tsuge, S., Ren, F., Kuroiwa, S., "Robust distributed speaker rocognition to solve the differences in frequency characteristic problem" *Proc. INFORMATION*, pp.207–210, Sep. 2004.

[4] Broun, C. C., Campbell, W. M., Pearce, D., Kelleher, H., "Distributed speaker recognition using the ETSI distributed Speech Recognition Standard," *A Speaker Odyssey - The Speaker Recognition Workshop*, pp.121–124, June 2001.

[5] Sit, C.-H., Mak, M.-W., and Kung, S.-Y., "Maximum Likelihood and Maximum A Posteriori Adaptation for Distributed Speaker Recognition Systems," *Proc. ICBA2004*, 2004.

[6] Grassi, S., Ansorge, M., Pellandini, F., Farine, P.-A., "Distributed speaker recognition using the ETSI AURORA standard," *Proc. 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, pp.120-125, Oct. 2002.

[7] Hirish, H.G., Pearce, D., "The AURORA experimental framework for the performance evaluations of speech recognition systems", *Proc. ASR2000*, Sep. 2000.

[8] Rubner, Y., Guibas, L., Tomasi, C., "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval," *Proc. the ARPA Image Understanding Workshop*, pp.661-668, May 1997.

[9] Umeda, Y., Tsuge, S., Ren, F., Kuroiwa, S., "Distributed Speaker Recognition using Earth Mover's Distance", *Proc. ICSLP2004*, pp.2389–2493, Oct. 2004.