

Distributed Speaker Recognition using Earth Mover's Distance

Yoshiyuki UMEDA Satoru TSUGE Fuji REN Shingo KUROIWA

Department of Information Science & Intelligent Systems
Faculty of Engineering, The University of Tokushima, Japan
{umeda,tsuge,ren,kuroiwa}@is.tokushima-u.ac.jp

Abstract

In this paper, we focus on distributed speaker recognition, a technique in which quantized feature parameters are sent to a server, as with distributed speech recognition. The Gaussian mixture model, the traditional method used for speaker recognition, is trained using the maximum likelihood approach. The GMM has output probability functions with continuous density functions. It is difficult to fit continuous density functions to quantized data. To overcome this problem, we propose a novel speaker recognition technique which does not need speaker model training. The proposed method directly calculates the distance between a set of quantized feature parameters of registered speech and a set of quantized feature parameters of test speech. To measure distance, we use Earth Mover's Distance (EMD). The EMD has recently been successfully applied to image retrieval. We conduct text-independent speaker identification experiments using the proposed method. When compared to results using the traditional GMM, the proposed method yielded relative error reductions of 80% for quantized data.

1. Introduction

In recent years, the use of the portable terminals, such as cellular phones and PDAs (Personal Digital Assistants), has become increasingly popular. Additionally, it is expected that almost all appliances will connect to the Internet in the future. As the result, it will be increasingly popular to control these appliances using mobile and hand-held devices. We believe that speaker recognition systems will be used for security. In conventional 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 is significantly degraded.

In speech recognition, a Distributed Speech Recognition (DSR) system is proposed to avoid these problems[1]. DSR separates the structural and computational components of recognition into two parts - front-end processing on the terminal and speech recognition 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, it is likely that speaker recognition will also shift to a distributed system. Some researchers have reported the results of distributed speaker recognition using ETSI DSR front-end[3, 4, 5].

The ETSI DSR front-end sends quantized feature parameters instead of the speech signal. These parameters are compressed to establish a lower bit rate for transmission. The ETSI DSR front-end employs a split vector quantization (VQ) algorithm for this compression algorithm. It is reported that the quantized data negatively affects recognition performances[3]. C.H. Sit *et al.*[5] reported that it is difficult to fit continuous density functions of GMM trained with M.L.A to quantized data. We can overcome this problem if the unquantized feature parameters are used to train the speaker model. However, we needed to prepare the collection method to obtain unquantized data. In this paper, we propose a novel distributed speaker recognition technique which does not require estimating parameter statistics of the speaker model. The proposed method directly calculates the distance between a set of quantized feature parameters of registered speech and a set of quantized feature parameters of test speech. To measure the distance, we tried to apply Earth Mover's Distance (EMD) to speaker recognition. EMD has been applied to calculate the distance between image data which were represented as histograms of multidimensional features[6]. Since the proposed method does not need to estimate statistical parameters, the proposed method is sufficient for the quantized data. In this paper, we report the results of experiments using the proposed method. Section 2 describes the perturbation of unquantized feature parameters. Section 3 explains the proposed speaker identification method, and section 4 presents our experiments for speaker identification.

2. Distributed Speaker Recognition

It is expected that the structure of a distributed speaker recognition system is similar to that of distributed speech recognition. Fig.1 illustrates the distributed speech

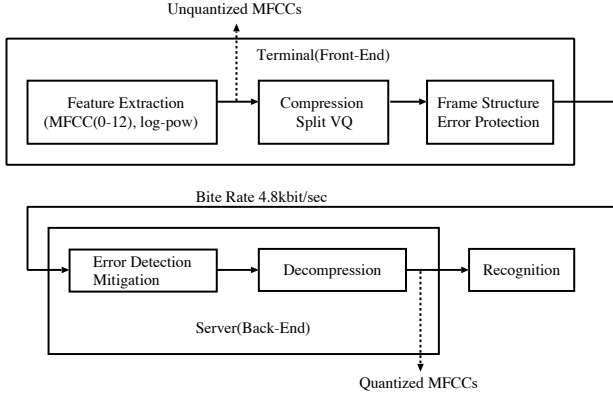


Figure 1: The block diagram of distributed speech recognition system

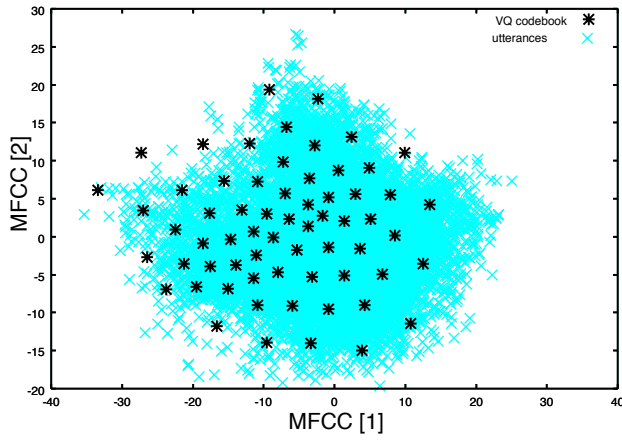


Figure 2: the scatter chart of the feature parameters (1st and 2nd order MFCCs). $\times$ is unquantized feature parameter. $*$ is quantized feature parameter.

recognition system. In this paper, we use the ETSI DSR front-end for the front-end of the distributed speaker recognition. This front-end compresses the feature parameters for transmission over the network. The ETSI DSR front-end employs vector quantization (VQ) for feature compression. Since the feature parameters on the server-side are quantized data, the distribution of the quantized data is discrete. Fig 2 shows the scatter chart of the feature parameters of 25 utterances, which are 1st and 2nd order MFCCs, and the centroid vectors in codebook employed by ETSI DSR front-end. We can see from this figure that many input vectors are quantized to one centroid and it is difficult to estimate the variance of continuous GMM. It is reported that quantized feature parameters negatively affect recognition performance[5]. Actually, we also found that some variances of GMM trained with the quantized feature parameters are adjusted

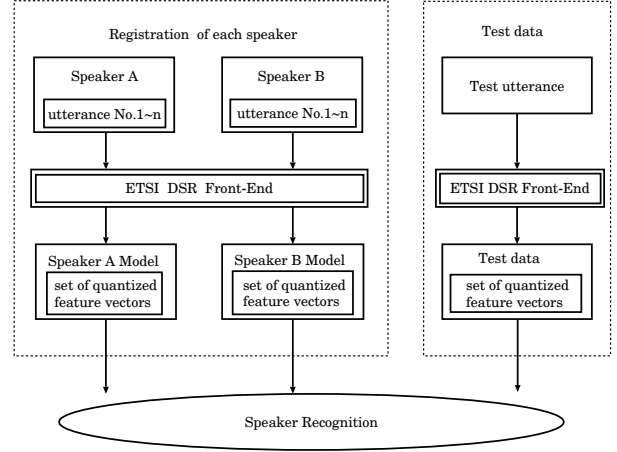


Figure 3: We describe the detail of the proposed method as following:

to the threshold (floored) since the continuous mixture model is estimated with discrete parameters. Furthermore, we observed that the variance flooring of the acoustic model increased the word error rate (WER) under the condition of distributed speech recognition[7]. Therefore, we consider it difficult to use the continuous GMM as a speaker model of distributed speaker recognition.

3. Proposed speaker recognition method using EMD

We described in section 2 that distributed speech recognition employs a VQ algorithm for feature compression so that speaker recognition should be performed using the quantized feature parameters. It has been reported in [3, 5] that the quantized data negatively affect the recognition performance. It is difficult to estimate the variance of continuous Gaussian mixture model because the quantized feature parameters are dispersed. Hence, we propose the non-parametric speaker recognition method using EMD. In the following section, we detail the proposed method.

3.1. The recognition flow of the proposed method

Figure 3 illustrates the flow of the proposed method.

- Registration of each speaker:
 1. Obtain the quantized feature parameters of each speaker. These feature parameters are extracted by ETSI DSR front-end.
 2. Register these feature parameters.
- Test data:
 1. Obtain quantized feature parameters extracted by ETSI DSR front-end of the test

speaker. These feature parameters used for the test data.

- Recognition:

1. Calculate the distance between each speaker model and the test data.
2. Identify the speaker whose model is closest to the test data.

In recognition, we apply the EMD to calculate the distance between the speaker model and test data. A brief overview of EMD is described in the following subsection.

3.2. Earth Mover's Distance

In this section, we overview the EMD algorithm. For full details, see [6]. The EMD is defined as the minimum amount of work needed to transform supplying points into demanding points. Hence, if we define the point as a feature parameters set, the EMD can be used to identify the distance between the speaker model and the test data.

The EMD computation has been formalized by the following linear programming problem[6]:

Given two points: $P = \{(p_1, w_{p_1}), \dots, (p_m, w_{p_m})\}$ and $Q = \{(q_1, w_{q_1}), \dots, (q_n, w_{q_n})\}$ where p_i and q_j are points in some Euclidean space, the MFCC vector space in our case, and w_{p_i}, w_{q_j} are the corresponding weights of the points, find an $m \times n$ cost matrix C where C_{ij} is the amount of weight of p_i matched to q_j , that will minimize the function:

$$\sum_{i=1}^m \sum_{j=1}^n C_{ij} \|p_i - q_j\| \quad (1)$$

($\|\cdot\|$ is the Euclidean distance)
subject to the following constraints:

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

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

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

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

The earth mover's distance is defined as the normalized

distance between points P and Q :

$$\begin{aligned} EMD(P, Q) &= \frac{\sum_{i=1}^m \sum_{j=1}^n C_{ij} \|p_i - q_j\|}{\sum_{i=1}^m \sum_{j=1}^n C_{ij}} \quad (6) \\ &= \frac{\sum_{i=1}^m \sum_{j=1}^n C_{ij} \|p_i - q_j\|}{\min(\sum_{i=1}^m w_{p_i}, \sum_{j=1}^n w_{q_j})} \quad (7) \end{aligned}$$

When we apply the EMD to speaker recognition, P and Q indicate the speaker model (the set of registered feature parameters) and the set of test feature parameters, respectively. In our distributed speaker recognition, weight is equivalent to the occurrence frequency of a corresponding quantized feature parameter.

4. Experiment

We conduct text independent speaker identification experiments to evaluate the proposed method.

4.1. Experimental conditions

The Japanese speech corpus we used supports making speaker models and test data of speaker identification. There were 21 male speakers; for each speaker, there were 7 sessions of 10 utterances each, for a total of 1,470 utterances. Each utterance is about 5 seconds in length. The 7 sessions were 3 months apart and started in August 1990. The speech signal was sampled at 16 kHz. We then down-sampled the speech to 8 kHz in order to investigate the effect of utterance sampling frequency on the identification error rate. We used the first 5 utterances of all speakers in the first session ($21 \times 5 = 105$ utterances) as speaker model data. The remaining 6 session utterances of the 21 speakers ($21 \times 10 \times 6 = 1,260$ utterances) were used for testing.

The 8 kHz and 16 kHz speech signals were segmented into overlapping frames of 25 ms, producing a frame every 10 ms. A Hamming window was applied to each frame. Mel-filtering was performed to extract 12 static MFCC (Mel-Frequency Cepstral Coefficients), as well as a logarithmic energy measure in the DSR front-end. 12 delta MFCC were created from the 12 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 + log-power).

In order to examine the effects of vector quantization, we used quantized feature parameters (*quantized*) and unquantized feature parameters (*unquantized*). CMS was performed on all the feature parameters. We also used the GMM with 16 mixtures which was trained with the same feature parameters using the registered data.

Table 1: The identification error rate at 16 kHz(in %)

feature vector(MFCCs)	method	IER in %
<i>unquantized</i>	GMM	2.0
<i>quantized</i>	GMM	2.5
<i>quantized</i>	EMD	0.5

Table 2: The identification error rate at 8 kHz(in %)

feature vector(MFCCs)	method	IER in %
<i>unquantized</i>	GMM	3.4
<i>quantized</i>	GMM	19.8
<i>quantized</i>	EMD	3.9

4.2. Experimental Results

Tables 1 and 2 show the speaker identification error rate (IER) obtained by using the proposed method (named EMD in this section) at sampling frequencies of 16 kHz and 8 kHz, respectively. We also show the IER using GMM, in these tables.

In Table 1 and 2, the GMM using quantized feature parameters increases the error rate compared with the GMM using unquantized data. When we investigated the speaker model of GMM which was trained with quantized feature parameters at 8kHz, we observed that many variance elements were flooded. This investigation shows that it is difficult to train the continuous GMM with quantized data.

On the other hand, the EMD using quantized data shows a much lower error rate than the GMM using *unquantized*. Actually, the EMD reduced errors 80% for quantized data.

In addition, the EMD using *quantized* shows the lower error rate than the GMM using *unquantized* of 16 kHz. In 8 kHz, the error rate of EMD using *quantized* is essentially the same as that of the error rate of GMM using *unquantized*. Although we proposed the EMD based method to increase robustness to the quantized parameters, it may have great potential for speaker recognition generally. We are now investigating the reason.

5. Summary

In this paper, we have proposed a novel speaker recognition method based on a non-parametric model using Earth Mover's Distance (EMD) for distributed speaker recognition. The EMD can directly calculate the distance between a set of quantized feature parameters. Experimental results on text-independent speaker identification showed that the proposed method give consistently better performance than the conventional method using GMM. We also obtained significant improvement in speaker identification performance (80.0% (at 16kHz) and 80.3% (at 8kHz)) with the proposed method.

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 Scientific Research (B), 14350204, 14380166, 2004, Hosono-Bunka Foundation (HBF) and International Communications Foundation (ICF).

7. References

- [1] B. Lilly and K. Paliwal, "Effect of speech coders on speech recognition performance," *ICSLP*, pp. 2344–2347, 1996.
- [2] "ETSI ES 201 108 v1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithms," 2000.
- [3] I. Fukuda, S. Tsuge, F. Ren, and S. Kuroiwa, "A speaker recognition using the ETSI standard DSR front-end," *ASJ*, pp. 175–176, Mar 2004.
- [4] C. C. Broun, W. M. Campbell, D. Pearce, and H. Kelleher, "Distributed speaker recognition using the ETSI distributed speech recognition standard," *A speaker Odyssey - The Speaker Recognition Workshop*, pp. 121–124, Jun 2001.
- [5] C.-H. Sit, M.-W. Mak, and S.-Y. Kung, "Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems," *ICBA*, 2004.
- [6] Y. Rubner, L. Guibas, and C. Tomasi, "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval," *In Proc. of the ARPA Image Understanding Workshop*, pp. 661–668, May 1997.
- [7] S. Kuroiwa and S. Tsuge, "Blind equalization techniques for ETSI standard DSR front-end," *icassp*, pp. 392–395, Apr 2003.