Fuzzy Cluster Analysis and its Evaluation Method

Shingo Kuroiwa ¹, Satoru Tsuge ² and Fuji Ren ³*

¹ Graduate School of Advanced Integration Science, Chiba University
² Institute of Technology and Science, The University of Tokushima
³ School of Information Engineering, Beijing University of Posts and Telecommunications

(received January 2, 2008)

Abstract: Recently, Distributed Speech Recognition (DSR) systems are widely deployed in Japanese cellular telephone networks. In these systems, personal authentication with voice is strongly desired. In this paper, we present several speaker recognition techniques developed in the University of Tokushima for Distributed Speaker Identification/Verification (DSI/DSV) systems. Especially, we present recent progress on a non-parametric speaker recognition system that is robust to quantization in the distributed systems comparing with conventional speaker recognition systems based on Gaussian Mixture Model (GMM). Evaluation results using the Japanese de facto standard speaker recognition corpus and CCC Speaker Recognition Evaluation 2006 data developed by the Chinese Corpus Consortium (CCC) show higher performance of the proposed method than GMM and VQ-distortion in the European Telecommunications Standards Institute (ETSI) DSR standard environment.

Keywords: Voice Authentication, Speaker Recognition, DSR, DSV, Earth Mover’s Distance

1 Introduction

In recent years, the use of portable terminals, such as mobile phones and PDAs (Personal Digital Assistants), has become increasingly popular. For such small devices, speech recognition is one of the most appropriate interfaces and is widely deployed in Japanese cellular telephone services. In these services, personal authentication with voice is strongly desired.

In order to meet this demand, we have proposed some speaker recognition techniques [1, 2, 3] that focused on Distributed Speech / Speaker Recognition (DSR) systems [4, 5, 6, 7, 8, 9, 10]. DSR separates the structural and computational components of recognition into two components - the front-end processing on the terminal and the matching block of the speech/speaker recognition on the server. One advantage of DSR is that it can avoid the negative effects of a speech codec because the terminal sends the server not a compressed speech signal but quantized feature parameters. Therefore, DSR can lead to an improvement in recognition performance. In speech recognition services, DSR is widely deployed in Japanese cellular telephone networks [11].

On the other hand, in speaker recognition, since a speaker model has to be trained with small amount of voice registration samples, quantization poses a big problem, especially in the case of using a continu-
recognition system. Fig. 1 shows a block diagram of the distributed speaker recognition. We use ETSI ES 201 108 v1.1.2[9] as the front-end of the system.

In DSR, many researchers use the 8kHz standard DSR front-end for speech recognition. Nevertheless, in speaker recognition, we should use higher sampling frequencies to improve recognition accuracy[12]. Accordingly, in this section, we try to compare the 8kHz standard and the 16kHz standard, whose transmitting bit rates are the same, 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 VQ. 64 centroids are used for each pair of cepstrum coefficients from 1st to 12th MFCC and 256 centroids are used for C0 and log energy. Since the feature vectors at the server-side are quantized data, the distribution of the quantized data becomes discrete.

In the next section, we investigate the influence of the quantized feature vectors on speaker recognition performance when using the parametric speaker model, GMM.

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 4.1.

Table 1 shows the experimental results using Japanese speaker recognition database described in 4.1, in which the best result for each condition is underlined. From this table, we can see that the performance of quantized feature vectors extracted from 16kHz sampling speech is better than 8kHz sampling speech even though they were compressed at the same bitrate.

Comparing the performance of the quantized feature vector with that of the unquantized feature vector in the table, we can observe that the quantization of feature vectors negatively affects the recognition performance expect when using 4 mixtures. 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 observed that many variance elements were floored when we investigated the speaker models that obtained the minimum IER. When we used speaker models with a smaller number of mixtures, the IER increased as shown in the table. Hence, it is difficult to estimate the variance, which is a statistical parameter, using the quantized feature vector.

3 Nonparametric speaker recognition method using EMD

In this section, we first provide a brief overview of Earth Mover's Distance (EMD)[13]. EMD directly calculates the distance between data sets, while GMM only calculates the likelihood of each datum. We think that the distribution of feature vectors in one utterance contains important information for speaker recognition, so we tried to use EMD in this paper. Next, we describe our proposed distributed speaker recognition method using a nonparametric speaker model and EMD measurement[2]. Finally, we propose a method to identify out-of-set data, which can be used for speaker verification system.

3.1 Earth Mover's Distance

The EMD was proposed by Rubner et al.[13] for an efficient image retrieval method. In this section, we describe the EMD algorithm.

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 $P = \{(p_1, w_1), \ldots, (p_m, w_m)\}$ be the
S. Kuroiwa, et al.: Distributes Speaker Identification and Variﬁcation

Figure 2: A block diagram of the feature extraction process and the proposed speaker recognition method[2]

discrete distribution, such as a histogram, where \( p_i \) is the centroid of each cluster and \( w_{pi} \) is the corresponding weight (= frequency) of the cluster; let \( Q = \{ (q_1, w_{q1}), \ldots, (q_m, w_{qm}) \} \) be the histogram of test feature vectors; and \( D = [d_{ij}] \) be the ground distance matrix where \( d_{ij} \) is the ground distance between centroids \( p_i \) and \( q_j \).

We want to ﬁnd a ﬂow \( F = [f_{ij}] \), with \( f_{ij} \) the ﬂow between \( p_i \) and \( q_j \) (i.e. the number of goods sent from \( p_i \) to \( q_j \), that minimizes the overall cost

\[
\text{WORK}(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij},
\]

subject to the following constraints:

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

\[
\sum_{j=1}^{n} f_{ij} \leq w_{pi} \quad (1 \leq i \leq m), \quad (3)
\]

\[
\sum_{i=1}^{m} f_{ij} \leq w_{qj} \quad (1 \leq j \leq n), \quad (4)
\]

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min \left( \sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj} \right). \quad (5)
\]

Constraint (2) allows moving goods from \( P \) to \( Q \) and not vice versa. Constraint (3) limits the amount of goods that can be sent by the cluster in \( P \) to their
weights. Constraint (4) limits the amount of goods that can be received by the cluster in $Q$ to their weights. Constraint (5) forces movement of 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 $F$, the EMD is defined as the work normalized by the total flow:

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

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

3.2 The recognition flow of the proposed method[2]

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 suppliers and the testing feature vectors as the consumers, the EMD can be applied to speaker recognition. Hence, we propose a distributed speaker recognition method using a nonparametric speaker model and EMD measurement. The proposed method represents the speaker model and testing feature vectors as a histogram. The detail of the proposed method is described as follows:

Fig. 2 illustrates a flow of the feature extraction process using the ETSI DSR standard and the proposed method. In the figure, dotted (‘) elements indicate data quantized once and double dotted (”) elements indicate data quantized twice. As shown in the upper part of the figure, both registered utterances and testing utterances are converted to quantized feature vector sequences, $V_A, V_B, \ldots$, and $V_X$, using the ETSI DSR front-end and back-end ($N_A, N_B$, and $N_X$ are the number of frames in each sequence). In this block, $c_t$ is a feature vector of time frame $t$ that consists of MFCC and logarithmic energy; $x_t$ is a code vector that is sent to the back-end (server); $\hat{c}_t$ is a decompressed feature vector; and $\tilde{v}_t$ is a feature vector for use in the subsequent speaker recognition process. Using $\hat{V}_A, \hat{V}_B, \ldots$, and $\hat{V}_X$, the proposed method is executed as follows:

(a) Speaker Model Generation Using the registered feature vectors, the system generates each speaker VQ codebook, $\{\tilde{p}_{1}^{sp}, \ldots, \tilde{p}_{m}^{sp}\}$, by using LBG algorithm with Euclidean distance, where $sp$ is a speaker name, and $m$ is a codebook size. In order to make a histogram of VQ centroids, the number of registered vectors whose nearest centroid is $\tilde{p}_{i}^{sp}$ is counted up and the number of the frequency is set to $w_{i}^{sp}$. As a result, we get a histogram of the speaker, $sp$, that is the speaker model in the proposed method,

$$P^{sp} = \{(\tilde{p}_{1}^{sp}, w_{1}^{sp}), \ldots, (\tilde{p}_{m}^{sp}, w_{m}^{sp})\}. \hspace{1cm} (7)$$

This histogram is used as the suppliers’ discrete distribution, $P$, described in the previous section.

(b) Testing data A histogram of testing data is directly calculated from $V_x$ that was quantized by ETSI DSR standard. The quantized feature vectors consist of static cepstrum vectors and their delta cepstrum vectors. Each static cepstrum vector is represented with six centroids of six independent codebooks whose sizes are 64, so that the quantized feature vectors have a huge number of possible combinations. Hence, we consider them as a set of vectors, $\{q_{1}^{X}, \ldots, q_{m_x}^{X}\}$, in which $m_x$ is the number of individual vectors. In order to make a histogram of the set of vectors, occurrence frequency of the vector $q_{i}^{X}$ is set to $w_{i}^{X}$. As a result, we get a histogram of the testing data,

$$Q^{X} = \{(q_{1}^{X}, w_{1}^{X}), \ldots, (q_{m_x}^{X}, w_{m_x}^{X})\}. \hspace{1cm} (8)$$

This histogram is used as the consumers’ discrete distribution, $Q$, described in the previous section.

(c) Identification Using the speaker models, $P^{sp}$, and the testing data, $Q^{X}$, speaker recognition is executed as in the following equation.

$$Speaker = \arg \min_{sp} EMD(P^{sp}, Q^{X}) \hspace{1cm} (9)$$

As the grand distance, $d_{ij}$, in EMD, we use the Euclidean distance between $\tilde{p}_{i}^{sp}$ and $q_{j}^{X}$. Since we utilize $w_{i}^{sp}$ and $w_{j}^{X}$ as the frequency of $\tilde{p}_{i}^{sp}$ and $q_{j}^{X}$, respectively, $f_{ij}$ is the number of vectors matched with $\tilde{p}_{i}^{sp}$ and $q_{j}^{X}$ (i.e. the number of goods sent from $\tilde{p}_{i}^{sp}$ to $q_{j}^{X}$), that minimizes the overall cost by EMD.

3.3 Identification of out-of-set data

In order to identify out-of-set data, we introduce an out-of-set identification module after “Speaker identification using EMD” in figure 2. Generally, open-set speaker identification evaluations include a candidate speaker list for each testing datum and speaker verification has one candidate speaker i.e. the claimant speaker. However, we calculate EMD between the testing datum and all speaker models. As the result,
an $N$-best ($N$ nearest) speakers list is obtained. After that, the $N$-best speaker list is compared with the provided candidate speaker list. If there is not a common speaker in the lists, the testing datum is rejected. On the other hand, if several speakers appear in the common speaker list, then the nearest speaker is chosen. $N$ is a parameter that controls False Rejection Rate (FRR) and False Acceptance Rate (FAR) in the method.

4 Experiments

First, we show text-independent speaker identification experiments using the Japanese de facto standard speaker recognition corpus. Next, we present our text-independent speaker identification evaluation result of the CCC Speaker Recognition Evaluation 2006 contest in ISCSLP2006. Finally, we conducted speaker verification experiments using the CCC Speaker Recognition Evaluation 2006 data.

4.1 Experiments using Japanese database

we conducted text-independent speaker identification experiments using a Japanese de facto standard speaker recognition corpus. From the corpus, we used 21 male speakers' utterances that were recorded in 7 sessions over 19 months. Each speaker spoke ten sentences, each of which had a length of about five seconds. For the registered data, i.e., the speaker model training data, we used five sentences which were uttered in the first session by each speaker. The utterances of the remaining six sessions were used for testing, in total 630 utterances (21 speakers $\times$ 5 sentences $\times$ 6 sessions). The text of these utterances was not contained in the training data.

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. The 12-dimensional MFCC and the logarithmic energy (log-energy) are calculated by the ETSI DSR front-end. After that the 12-dimensional delta-MFCC and delta-logarithmic energy were extracted from the quantized MFCC and the logarithmic energy to constitute a 25-dimensional feature vector (12 static MFCCs $+$ 12 delta MFCC $+$ delta log-energy). Cepstrum Mean Subtraction (CMS) [14] was applied on the static MFCC vectors.

For comparison with the proposed method, we also conducted experiments with speaker recognition methods based on GMM[15] and VQ-distortion[16].

Table 2 shows the experimental results. We used the ETSI DSR standard for feature extraction, but we skipped the quantization process in the case of “Un-quantized”. These results show that the proposed method is an effective method for not only “Quantized” data but also “Un-quantized” data. From the table, we can see the proposed method gave the better performance than the conventional methods, GMM and VQ-distortion, even when they used un-quantized data.

4.2 CCC Speaker Recognition Evaluation

To evaluate the proposed method using a larger database, we have taken the challenge of CCC Speaker Recognition Evaluation 2006[17] organized by the Chinese Corpus Consortium (CCC) for the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP 2006). Although the proposed method was designed for the DSR environment, we show it achieves higher performance than the conventional methods even in the conventional telephone environments.

4.2.1 Evaluation Data

CCC provided several kinds of tasks, i.e., text-independent speaker identification, text-dependent and text-independent speaker verification, text-independent cross-channel speaker identification, and text-dependent and text-independent cross-channel speaker verification. We participated the text-independent speaker recognition task in the contest in view of the characteristics of the proposed method. In this paper, we have additionally conducted experiments on the text-independent speaker verification task.

The data set of the speaker identification task contained 400 speakers' data for enrollment, and 2,395 utterances for testing. Each datum to enroll is longer than 30 seconds and recorded over a land-line (PSTN) or cellular-phone (GSM only) network. The channel each speaker used to speak the utterances was the same across enrollment and testing data. The recording of each speaker was conducted in a single session, so that there was no time difference between enrollment and testing. Each testing datum has a candidate speakers list and about half of the testing data

<table>
<thead>
<tr>
<th>Method</th>
<th>Un-quantized</th>
<th>Quantized</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>1.6%</td>
<td>4.0%</td>
</tr>
<tr>
<td>VQ-distortion</td>
<td>0.8%</td>
<td>1.0%</td>
</tr>
<tr>
<td>EMD (proposed)</td>
<td></td>
<td>0.6%</td>
</tr>
</tbody>
</table>
were uttered by out-of-set speakers who did not appear in the list. Therefore, the speaker identification algorithm has to decide whether each testing datum is in-set or out-of-set also.

On the other hand, the data set of speaker verification task contained 800 speakers’ data for enrollment. The length of each data is from 21 sec. to 85 sec. (average 37 sec.) Testing data consisted of 347 true-speakers and 5,586 imposters. The length of each data is from 4.8 sec. to 54 sec. (average 16 sec.) Recording conditions were the same as the speaker identification task.

CCC also provided development data that contained 300 speakers’ utterances with speaker labels and channel conditions. We were able to decide the various parameters of the algorithm using the development data.

The performance of speaker identification was evaluated by Identification Correctness Rate (ICR), defined as:

\[
\text{%ICR} = \frac{\text{Number Of Correctly Identified Data}}{\text{Total Number of Trial Data}} \times 100\%.
\]

where “correctly identified data” means those data identified as the speaker models they should be by the top-candidate output, if they were “in-set”, or “non-match” if “out-of-set”.

4.2.2 Experimental conditions

All data, sampled at 8kHz, were segmented into overlapping frames of 25ms, producing a frame every 10ms. We constituted a 25-dimensional feature vector (12 static MFCCs + 12 delta MFCCs + delta log-energy) by the same way as the previous section.

In the experiment, we set the number of centroids of each speaker’s codebook to 64 and set the parameter \(N\) for out-of-set speaker rejection to 4 for the speaker-identification task, which gave the best accuracy in experiments using the development data.

4.2.3 Experimental results of identification task

Table 3 shows the Identification Correctness Rate (ICR), False Acceptance Rate (FAR), False Rejection Rate (FRR), and Recognition Error Rate (RER). RER is the rate which identified the utterance of one speaker in the candidate list as another speaker’s utterance. The table shows the experimental result of the proposed method. This result is the best ICR in the “speaker identification task” under the closed-channel condition of CCC Speaker Recognition Evaluation 2006 in ISCSLP2006. This means that the proposed method achieved higher performance than the GMM-based techniques[17, 18] in this task. We expect the reason for this result is the difference of distance measures (score calculation). The proposed method directly calculates the distance between data sets, while GMM-based methods calculate the score by totaling the likelihood of each frame. The proposed method can compare the distribution of the speaker model with the distribution of the testing feature vectors. On the other hand, the proposed algorithm required a large computation time. Actually, it took about nine minutes to identify one utterance using an Intel Pentium 4 3.2GHz processor in the experiments.

When we investigated the data of FAR and FRR, the word sequences of several testing data were included in the training data of the other speaker and not included in the training data of the correct speaker. The use of automatic speech recognition for phoneme-dependent identification methods will improve the speaker identification performance for these data[1, 19], although it will turn into a language dependent system.

4.2.4 Experimental results of verification task

Figure 3 presents the detection error tradeoff (DET) curve of the proposed method. The plot was generated by varying the ranking parameter \(N\) described in 3.3. False acceptance rates and False rejection rates are also listed in Table 4. Comparing with the DET curves obtained by participants of the CCC Speaker Recognition Evaluation 2006 contest in ISCSLP2006, which include speaker verification methods using GMM, Support Vector Machine, and so on[17], our results achieved higher performance when FAR is less than 1.0 %.
5 Summary

In this paper, we have presented a non-parametric speaker identification and verification method using Earth Mover’s Distance (EMD) designed for the distributed systems. Experimental results using Japanese de facto standard speaker recognition corpus showed the proposed method achieved higher performance than the conventional methods using GMM and VQ-distortion in the ETSI DSR standard environment. The results for CCC Speaker Recognition Evaluation 2006 organized by the Chinese Corpus Consortium (CCC) for the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP 2006) also showed the proposed method is an effective speaker identification and verification method in the conventional telephone networks.

In future work, we will accelerate the distance calculation process in the proposed algorithm and evaluate it in Japanese cellular networks. Furthermore, we will consider use of speech recognition to improve the speaker identification accuracy and verification performance [1].

Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Exploratory Research, 17656128, Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B), 17300065, 19300029. The authors also greatly appreciate the Chinese Corpus Consortium who provided the valuable speech data.

References


