Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 239-254

Speaker Identification Method Using Earth Mover's Distance for CCC Speaker Recognition Evaluation 2006

Shingo Kuroiwa[*], Satoru Tsuge[*], Masahiko Kita[*], and Fuji Ren[*+]

Abstract

In this paper, we present a non-parametric speaker identification method using Earth Mover's Distance (EMD) designed for text-indepedent speaker identification and its evaluation results for *CCC Speaker Recognition Evaluation 2006*, organized by the Chinese Corpus Consortium (CCC) for the *th International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006). EMD based speaker identification (EMD-IR) was originally designed to be applied to a distributed speaker identification system, in which the feature vectors are compressed by vector quantization at a terminal and sent to a server that executes a pattern matching process. In this structure, we had to train speaker models using quantized data, then we utilized a non-parametric speaker model and EMD. From the experimental results on a Japanese speech corpus, EMD-IR showed higher robustness to the quantized data than the conventional GMM technique. Moreover, it achieved higher accuracy than GMM even if the data was not quantized. Hence, we have taken the challenge of *CCC Speaker Recognition Evaluation 2006* using EMD-IR. Since the identification tasks defined in the evaluation were on an open-set basis, we introduce a new speaker verification module. Evaluation results show that EMD-IR achieves 99.3 % *Identification Correctness Rate* in a closed-channel speaker identification task.

Keywords: Speaker Identification, Earth Mover's Distance, Non-Parametric, Vector Quantization, Chinese Speech Corpus

[*] Institute of Technology and Science, The University of Tokushima, 2-1 Minami-Josanjima, Tokushima-shi 770-8506, Japan Tel: +81 886569689 Fax: +81 886560575
E-mail: kuroiwa@is.tokushima-u.ac.jp

[+] School of Information Engineering, Beijing University of Posts and Telecommunications Beijing 100876, China

[Received April 3, 2007; Revised August 17, 2007; Accepted August 20, 2007]

1. Introduction

In recent years, the use of portable terminals, such as mobile 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 a result, it will become increasingly popular to control these appliances using mobile and hand-held devices. We believe that a speaker recognition system will be used as a convenient personal identification system in this case.

In order to meet this demand, we have proposed some speaker recognition techniques [Fattah 2006A; Kuroiwa 2006; Fattah 2006B] that have focused on Distributed Speech/Speaker Recognition (DSR) systems [Pearce 2000; Broun 2001; Grassi 2002; Sit 2004; Fukuda 2004; ETSI 2000; ITU 2004]. 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 quantized feature parameters instead of a compressed speech signal. Therefore, DSR can lead to an improvement in recognition performance. DSR is widely deployed in Japanese cellular telephone networks for speech recognition services [KDDI 2006]. On the other hand, in speaker recognition, since a speaker model has to be trained with a small amount of voice registration samples, quantization poses a big problem, especially in the case of using a continuous probability density function, *e.g.* GMM [Sit 2004; Fukuda 2004].

To solve this problem, we proposed a non-parametric speaker recognition method that does not require previous assumption of any probability distribution function and estimation of statistical parameters such as mean and variance for the speaker model [Kuroiwa 2006]. We represented a speaker model using a histogram of speaker-dependent VQ codebooks (VQ histogram). To calculate the distance between the speaker model and the feature vectors for recognition, we applied the Earth Mover's Distance (EMD) algorithm. The EMD algorithm has been applied to calculate the distance between two images represented by histograms[1] of multidimensional features [Rubner 1997]. In Kuroiwa [2006], we conducted text-independent speaker identification experiments using the Japanese *de facto* standard speaker recognition corpus and obtained better performance than GMM for quantized data. After that, we extended the algorithm to calculate the distance between a VQ histogram and a data set. From the results, we observed it achieved higher accuracy than the GMM and VQ distortion methods even if the data was not quantized. We believe that the better results were obtained by the proposed method because it considers not only the centroid location, but also the weight.

[1] In Rubner [1997], EMD is defined as the distance between two *signatures*. The *signatures* are histograms that have different bins, to that effect we use the term "histogram" in this paper.

EMD can compare the distribution of the speaker model with the distribution of the testing feature vectors as is.

To evaluate the proposed method using a larger database, we have taken the challenge of *CCC Speaker Recognition Evaluation 2006* [Zheng 2006] organized by the Chinese Corpus Consortium (CCC) for *the 5th International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006). In view of the characteristics of the proposed method, we have chosen the text-independent speaker recognition task from the five tasks in *CCC Speaker Recognition Evaluation 2006*. The method was originally designed for the classic speaker identification problem that does not require a function to reject out-of-set speaker voices. However, since the evaluation data includes out-of-set speaker voices, we introduce a new speaker verification module in this paper. We also introduce a voice activity detector that classifies each frame as either a valid speech frame or a nonvalid frame (background noise or unreliable speech) on a frame-by-frame basis, in order to avoid miss-identification caused by non-speech frame information.

This paper will continue as follows. Section 2 explains the Earth Mover's Distance and the originally proposed speaker identification method. Some modifications for *CCC Speaker Recognition Evaluation 2006* and its evaluation results for the Japanese *de facto* standard speaker recognition corpus are also described. Section 3 presents speaker identification experiments using *CCC Speaker Recognition Evaluation* corpus. Finally, we summarize this paper in Section 4.

2. Non-Parametric Speaker Recognition Method Using EMD

In this section, we first provide a brief overview of Earth Mover's Distance. Next, we describe the distributed speaker recognition method using a non-parametric speaker model and EMD measurement. Finally, we propose EMD speaker identification for non-quantized data and a speaker verification module for identifying out-of-set speaker voices.

2.1 Earth Mover's Distance

EMD was proposed by Rubner [1997] as an efficient image retrieval method. In this section, we describe the EMD algorithm.

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 $\boldsymbol{P} = \{(\boldsymbol{p}_1, w_{p1}), \ldots, (\boldsymbol{p}_m, w_{pm})\}$ 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 $\boldsymbol{Q} = \{(\boldsymbol{q}_1, w_{q1}), \ldots, (\boldsymbol{q}_n, w_{qn})\}$ 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 $\boldsymbol{F} = [f_{ij}]$, where f_{ij} is the flow between $\boldsymbol{p}_i$ and $\boldsymbol{q}_j$ (*i.e.* the number of *goods* sent from $\boldsymbol{p}_i$ to $\boldsymbol{q}_j$), that minimizes the overall cost:

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

subject to the following constraints

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

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

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

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

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 clusters in $\boldsymbol{Q}$ 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 $\boldsymbol{F}$, the EMD is defined as the work normalized by the total flow:

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

The normalization factor is the total weight of a smaller distribution, due to 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. In order to find the optimal flow, we used “EMD.c”, which has been made by available by Rubner [1999], in the following experiments. This program uses the transportation-simplex method and its computational complexity increases exponentially with the number of histogram bins [Rubner 1997].

2.2 Recognition Flow of the Proposed Method

In the previous section, we described the concept 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 non-parametric speaker model and EMD measurement.

The proposed method represents the speaker model and testing feature vectors as histograms. The details of the proposed method are described as follows.

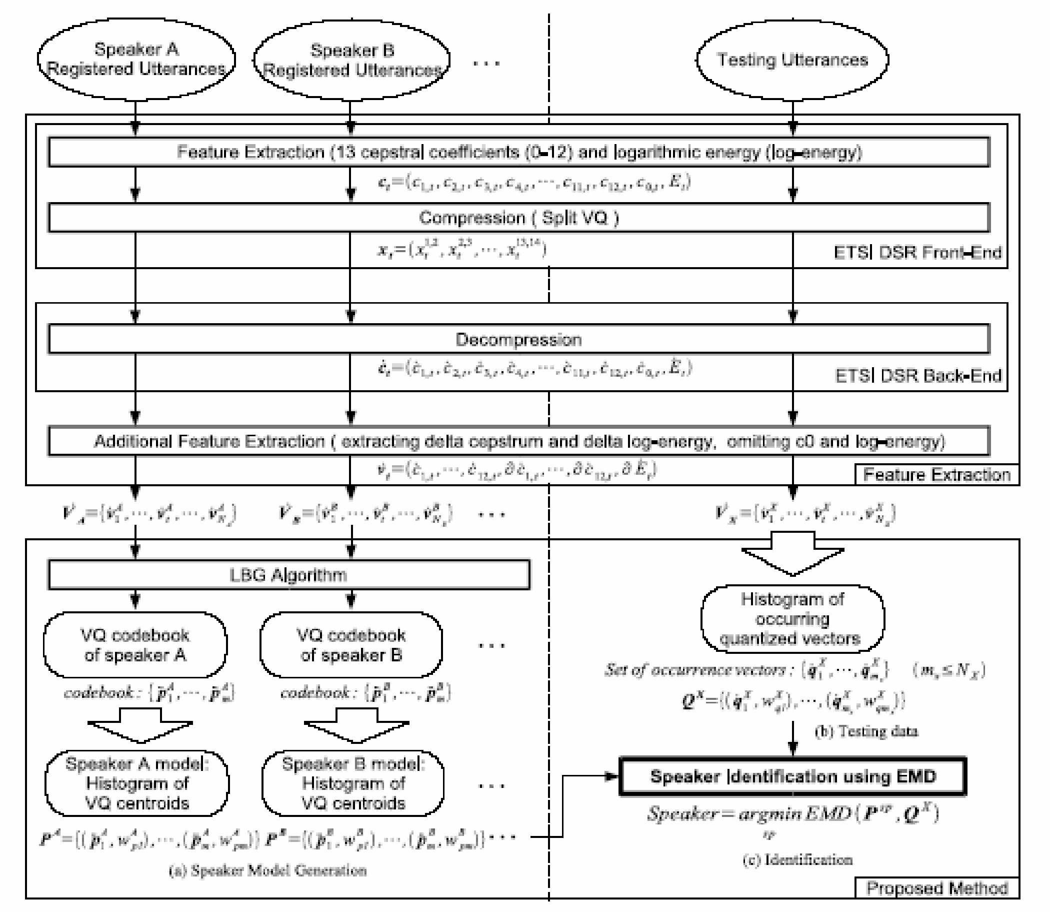
Speaker A Registered Utterances
Speaker B Registered Utterances
. . .
Testing Utterances
Feature Extraction (13 cepstral coefficients (0-12) and logarithmic energy (log-energy)
$c_t=(c_{1,t},c_{2,t},c_{3,t},c_{4,t},\cdots,c_{11,t},c_{12,t},c_{0,t},E_t)$
Compression (Split VQ)
$x_t=(x_t^{1,2},x_t^{2,3},\cdots,x_t^{13,14})$
ETSI DSR Front-End
Decompression
$\dot{c}_t=(\dot{c}_{1,t},\dot{c}_{2,t},\dot{c}_{3,t},\dot{c}_{4,t},\cdots,\dot{c}_{11,t},\dot{c}_{12,t},\dot{c}_{0,t},\dot{E}_t)$
ETSI DSR Back-End
Additional Feature Extraction (extracting delta cepstrum and delta log-energy, omitting c0 and log-energy)
$\dot{v}_t=(\dot{c}_{1,t},\cdots,\dot{c}_{12,t},\partial\dot{c}_{1,t},\cdots,\partial\dot{c}_{12,t},\partial\dot{E}_t)$
Feature Extraction
$\dot{V}_A=\{\dot{v}_1^A,\cdots,\dot{v}_t^A,\cdots,\dot{v}_{N_A}^A\}$
$\dot{V}_B=\{\dot{v}_1^B,\cdots,\dot{v}_t^B,\cdots,\dot{v}_{N_B}^B\}$
. . .
$\dot{V}_X=\{\dot{v}_1^X,\cdots,\dot{v}_t^X,\cdots,\dot{v}_{N_X}^X\}$
LBG Algorithm
VQ codebook of speaker A
VQ codebook of speaker B
. . .
codebook : $\{\ddot{p}_1^A,\cdots,\ddot{p}_m^A\}$
codebook : $\{\ddot{p}_1^B,\cdots,\ddot{p}_m^B\}$
Speaker A model: Histogram of VQ centroids
Speaker B model: Histogram of VQ centroids
. . .
$P^A=\{(\ddot{p}_1^A,w_{p1}^A),\cdots,(\ddot{p}_m^A,w_{pm}^A)\}$ $P^B=\{(\ddot{p}_1^B,w_{p1}^B),\cdots,(\ddot{p}_m^B,w_{pm}^B)\}$ · · ·
(a) Speaker Model Generation
Histogram of occurring quantized vectors
Set of occurrence vectors : $\{\dot{q}_1^X,\cdots,\dot{q}_{m_x}^X\}$ $(m_x \le N_X)$
$Q^X=\{(\dot{q}_1^X,w_{q1}^X),\cdots,(\dot{q}_{m_x}^X,w_{qm_x}^X)\}$
(b) Testing data
Speaker Identification using EMD
$Speaker=\underset{sp}{argmin}\,EMD(P^{sp},Q^X)$
(c) Identification
Proposed Method

Figure 1. A block diagram of the feature extraction process and the proposed speaker recognition method [Kuroiwa 2006]

Figure 1 illustrates the outline of the feature extraction process using the ETSI DSR standard [ETSI 2000] 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, $\dot{V}_A,\dot{V}_B,\ldots,$ and $\dot{V}_X$, 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, $\boldsymbol{c}_t$ is a feature vector of time frame t that consists of MFCC and logarithmic energy; $\boldsymbol{x}_t$ is a code vector that is sent to the back-end (server); $\dot{\boldsymbol{c}}_t$ is a decompressed feature vector; and $\dot{\boldsymbol{v}}_t$ is a feature vector for use in the subsequent speaker recognition process. Using $\dot{V}_A,\dot{V}_B,\ldots,$ and $\dot{V}_X$, the proposed method is executed as follows.

(a) Speaker Model Generation

Using the registered feature vectors, the system generates each speaker's VQ codebook, $\{\ddot{\boldsymbol{p}}_1^{sp},\ldots,\ddot{\boldsymbol{p}}_m^{sp}\}$, using the LBG algorithm with Euclidean distance where sp is the speaker name and m is the codebook size. In order to make a histogram of VQ centroids, the number of registered vectors whose nearest centroid is $\ddot{\boldsymbol{p}}_i^{sp}$ is counted and the frequency is set to $w_{p_i}^{sp}$.[2]

As a result, we get a histogram of the speaker, sp, that is used as the speaker model in the proposed method:

$$\boldsymbol{P}^{sp} = \{(\ddot{\boldsymbol{p}}_1^{sp}, w_{p_1}^{sp})\ldots,(\ddot{\boldsymbol{p}}_m^{sp}, w_{p_m}^{sp})\}\,. \tag{7}$$

This histogram is used as the *suppliers*' discrete distribution, $\boldsymbol{P}$, described in the previous section.

(b) Testing data

A histogram of the testing data is directly calculated from $\dot{\boldsymbol{V}}_X$, which was quantized by the ETSI DSR standard. The quantized feature vectors consist of static cepstrum vectors that have 64^6 possible combinations and their delta cepstrum vectors, creating a set of vectors, $\{\dot{\boldsymbol{q}}_1^X,\ldots,\dot{\boldsymbol{q}}_{m_x}^X\}$, where m_X is the number of individual vectors. In order to create a histogram from the set of vectors, the occurrence frequency of the vector $\dot{\boldsymbol{q}}_i^X$ is set to $w_{q_i}^X$. As a result, we get a histogram of the testing data:

$$\boldsymbol{Q}^X = \{(\dot{\boldsymbol{q}}_1^X, w_{q_1}^X)\ldots,(\dot{\boldsymbol{q}}_{m_x}^X, w_{q_{m_X}}^X)\}\,. \tag{8}$$

This histogram is used as the *consumers*' discrete distribution, $\boldsymbol{Q}$, described in the previous section.

(c) Identification

Using the speaker models, $\boldsymbol{P}^{sp}$, and the testing data, $\boldsymbol{Q}^X$, speaker recognition is executed as in the following equation:

$$Speaker = \underset{sp}{argmin}\, EMD(\boldsymbol{P}^{sp},\boldsymbol{Q}^X)\,. \tag{9}$$

For the ground distance d_{ij}, in EMD, we used the Euclidean distance between $\ddot{\boldsymbol{p}}_i^{sp}$ and $\dot{\boldsymbol{q}}_j^X$. Since the frequencies of $\ddot{\boldsymbol{p}}_i^{sp}$ and $\dot{\boldsymbol{q}}_j^X$ were used as $w_{p_i}^{sp}$ and $w_{q_j}^X$, f_{ij} is the number of matched vectors in $\ddot{\boldsymbol{p}}_i^{sp}$ and $\dot{\boldsymbol{q}}_j^X$ (*i.e.* the number of *goods* sent from $\ddot{\boldsymbol{p}}_i^{sp}$ to $\dot{\boldsymbol{q}}_j^X$) that minimizes the overall cost by EMD.

[2] Although EMD does not satisfy the "Commutative Property" without weight normalization, we used the raw frequency counts as the weight. This is because we assume that the registration speech is longer than the testing speech, that is, we expect a set of phoneme frames of the testing speech to be a subset of phoneme frames of the registration speech.

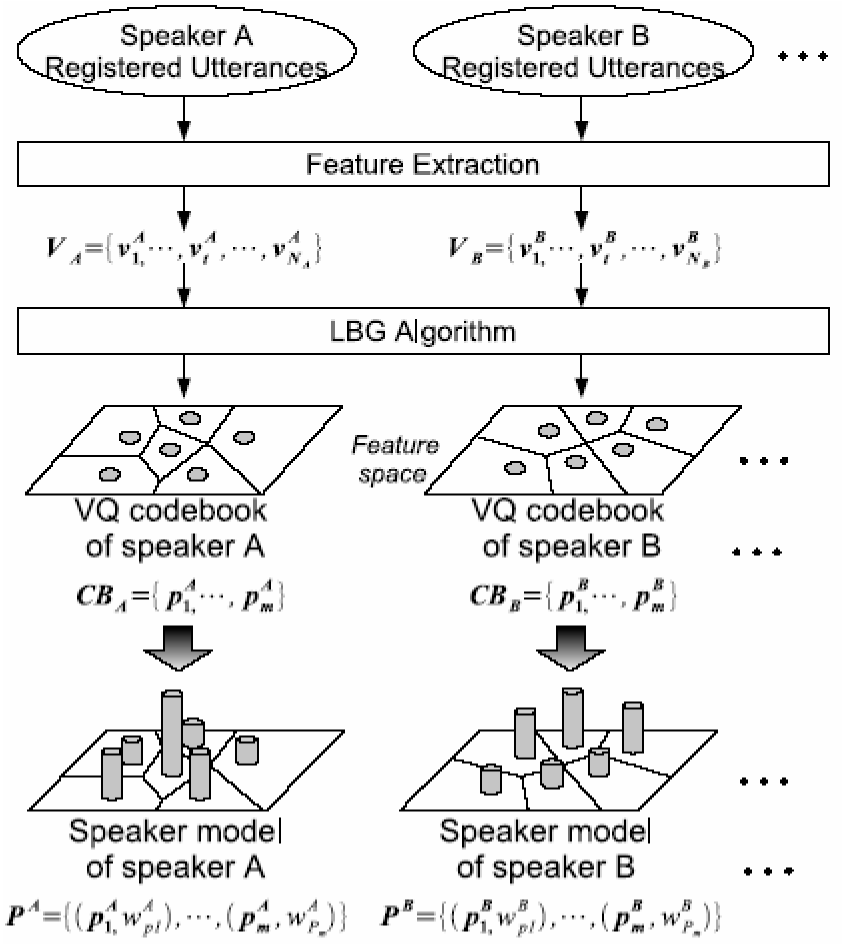
Speaker A Registered Utterances
Speaker B Registered Utterances
Feature Extraction
$V_A=\{v_1^A,\cdots,v_t^A,\cdots,v_{N_A}^A\}$
$V_B=\{v_1^B,\cdots,v_t^B,\cdots,v_{N_B}^B\}$
LBG Algorithm
Feature space
VQ codebook of speaker A
VQ codebook of speaker B
$CB_A=\{p_1^A,\cdots,p_m^A\}$
$CB_B=\{p_1^B,\cdots,p_m^B\}$
Speaker model of speaker A
Speaker model of speaker B
$P^A=\{(p_1^A,w_{p1}^A),\cdots,(p_m^A,w_{P_m}^A)\}$
$P^B=\{(p_1^B,w_{p1}^B),\cdots,(p_m^B,w_{P_m}^B)\}$

Figure 2. Block diagram of speaker model creation

2.3 Modifications for Non-Quantized Data

In order to apply the proposed method to non-quantized data, we have modified the recognition flow described in the previous section.

First, the "Compression" and "Decompression" blocks in Figure 1 are skipped, and consequently, feature vector sequences $\dot{V}_A, \dot{V}_{B,\ldots}$, and $\dot{V}_X$ become non-quantized feature vector sequences $V_A, V_B, \ldots$, and V_X . In "Speaker Model Generation", the LBG algorithm can generate each speaker's codebook from the non-quantized feature vector sequence without any modification of the algorithm. Figure 2 shows a block diagram of this speaker model creation process.

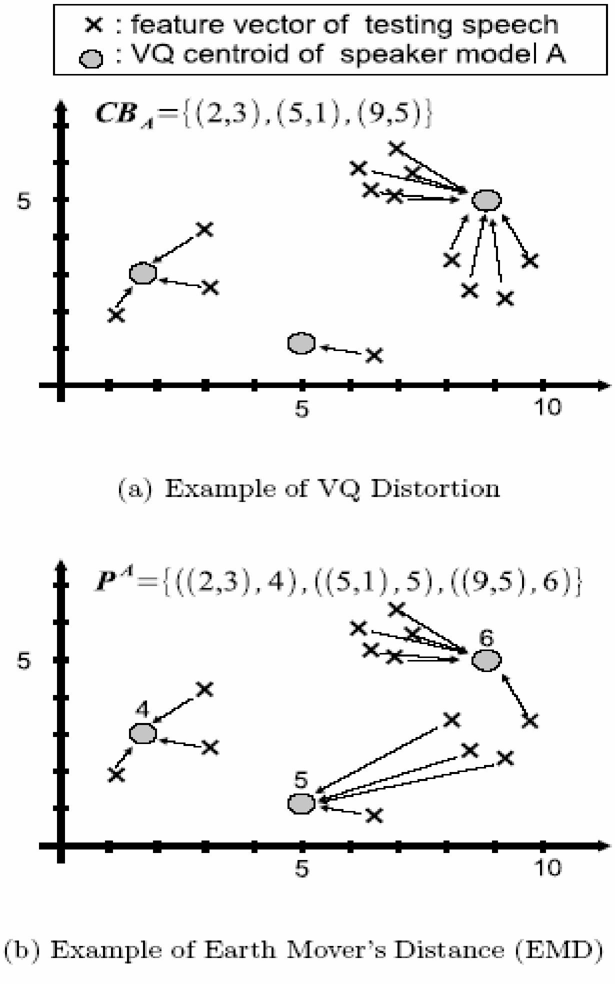
× : feature vector of testing speech
○ : VQ centroid of speaker model A
CB_A={(2,3),(5,1),(9,5)}
5
5
10
(a) Example of VQ Distortion
P^A={((2,3),4),((5,1),5),((9,5),6)}
6
4
5
5
5
10
(b) Example of Earth Mover's Distance (EMD)

Figure 3. Conceptual image of the difference of VQ and EMD

In the identification process, we consider the test utterance's set of the feature vectors to be a histogram in which the occurrence frequency of each vector is one. Figure 3 shows conceptual images of the speaker identification score calculation in the VQ distortion method and the proposed EMD method. The number written above each circle (centroid) in figure (b) is the weight or amount of data that each centroid can accept. The VQ distortion method does not care about the amount of data assigned to each centroid. This results in the VQ distortion becoming small when many vectors concentrate on a single centroid, which is caused by specific sounds, such as tone-like noises, the sound of breathing, etc. On the other hand, EMD takes into account the amount of data for each centroid. This means that the proposed method can compare the distribution of the speaker model with the distribution of the testing feature vectors.

Through above modification, we can calculate the EMD between the speaker model and the non-quantized testing data. To confirm the performance of this modification, we conducted text-independent speaker identification experiments using the 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 there were 630 utterances (21 speakers × 5 sentences × 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. Mel-filtering was performed to extract 12-dimensional static MFCC, as well as a logarithmic energy measure in the DSR front-end. The 12-dimensional delta MFCC was extracted from the static MFCC to constitute a 25-dimensional feature vector (12 static MFCCs + 12 delta MFCC + delta log-energy). Cepstral Mean Subtraction (CMS) [Atal 1974] was applied on the static MFCC vectors.

For comparison with the proposed method, we also conducted experiments with speaker recognition methods based on GMM [Reynolds 1995; Kuroiwa 2006] and VQ-distortion [Soong 1985; Kuroiwa 2006].

In the experiment, the number of centroids for each speaker's codebook was set to 256 for both the proposed method and the VQ-distortion based method. The GMM based method used a diagonal covariance with 64 components. These parameter settings obtained the best results [Kuroiwa 2006]. The LBG algorithm was used for training the VQ codebooks, and the Baum-Welch maximum likelihood algorithm was used for training the GMMs. HTK3.3 [Young 2005] was utilized for both of the training sets.

Table 1 shows the experimental results. We used the ETSI DSR standard for feature extraction, but we skipped the quantization process in the case of "non-quantized".

Table 1. Identification error rate for the Japanese database

Method	Non-quantized	Quantized
GMM	1.6 % (10/630)	4.0 % (25/630)
VQ-distortion	0.8 % (5/630)	1.0 % (6/630)
EMD (proposed)	0.6 % (4/630)	0.6 % (4/630)

These results show that the proposed method is an effective method for not only "Quantized" data but also "Non-quantized" data.

2.4 Identification of Out-of-Set Data

In order to identify out-of-set data, which is needed for the *CCC Speaker Recognition Evaluation* corpus, we introduce an out-of-set identification module after "Speaker identification using EMD" in Figure 1. The evaluation includes a candidate speaker list for each testing datum. However, we calculate the EMD between the testing datum and all speaker models. This results in an N-best (N nearest) speaker list being obtained. Then, the N-best speaker list is compared with the provided candidate speaker list. If no common speaker exists between 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. It is most likely dependent on the total number of speaker models. In the following experiments, we used 400 speaker models that were trained with all data for enrollment in the text-independent speaker recognition task of *CCC Speaker Recognition Evaluation 2006*. N was set to 4, which made the ratio of data for in-set and out-of-set about 1:1. This matched the previous information provided with the testing data. We think this is reasonable because, in a real system, we can obtain the utterances each speaker used to access the system and from this we can know the ratio of in-set and out-of-set users in a field trial phase of the system. Actually, we have a good example of this technique, the threshold values in the Prank Call Rejection System [Kuroiwa 1996], deployed by KDDI international telephone service from 1996, were determined with this kind of process which still works effectively today.

2.5 Voice Activity Detector

In order to avoid any detrimental effects caused by non-speech sections and unreliable speech frames, we employed a voice activity detector (VAD) that classifies each frame as either speech or background noise on a frame-by-frame basis. The VAD uses a power threshold that was calculated from percentile levels based on each observed speech signal. We used the following threshold in the experiments.

$$Threshold = (P_{95\%tile} - P_{10\%tile}) \times \alpha + P_{10\%tile}, \tag{10}$$

Only the frames with a higher power level than this threshold value were used for speaker identification.

α is set to 0.2, which allowed the proposed method to obtain a good identification correctness rate for the development data in *CCC Speaker Recognition Evaluation 2006*. This process reduced the number of frames by 10 % to 50 %. This reduction greatly benefits the proposed method, since it is computationally expensive.

3. Experiments

We conducted text-independent speaker identification experiments to evaluate the proposed method using the *CCC Speaker Recognition Evaluation 2006* data developed by the Chinese Corpus Consortium (CCC).

3.1 Task Definition

In *the 5th International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006), the CCC organized a special session on speaker recognition and provided speech data to evaluate speaker recognition algorithms using the same database. The CCC provided several kinds of tasks: 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 chose the text-independent speaker recognition task in view of the characteristics of the proposed method. The data set of this task contained 400 speakers' data for enrollment, and 2,395 utterances for testing. Each datum to enroll was 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. Each testing datum had a candidate speakers list, and about half of the testing data was uttered by out-of-set speakers who did not appear in the list. Therefore, the speaker identification algorithm had to decide whether each testing datum was in-set or out-of-set also.

The 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 using the *Identification Correctness Rate*, defined as:

$$\%CorrectIdentification = \frac{NumberOfCorrectlyIdentifiedData}{TotalNumberOfTrialData} \times 100\%, \tag{11}$$

where "correctly identified data" means the 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".

3.2 Experimental Conditions

All data, sampled at 8kHz, was 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 MFCC, as well as a logarithmic energy (log-energy) measure. The 12-dimensional delta MFCC and delta log-energy were extracted from the static

MFCC and the log-energy, respectively. After that, by omitting the log-energy, we constituted a 25-dimensional feature vector (12 static MFCCs + 12 delta MFCCs + delta log-energy). Cepstral Mean Subtraction (CMS) was applied on the static MFCC vectors. We used HTK3.3 [Young 2005] for feature extraction.

In the experiment, we set the number of centroids of each speaker's codebook to 64, which gave the best accuracy in experiments using the development data. The parameter for detecting the out-of-set data was also set up using this data along with the previous information that the ratio of testing samples for in-set and out-of-set cases would be about 1:1.

3.3 Experimental Results

Table 2 shows the *Identification Correctness Rate* (ICR), False Acceptance Rate (FAR), False Rejection Rate (FRR), and Recognition Error Rate (RER). RER is the rate in which one speaker's utterance was identified as another's in the candidate list. The table shows the proposed method achieved extremely high performance in the task. This result is the best ICR in the "speaker identification task" under the closed-channel condition of *CCC Speaker Recognition Evaluation 2006* in ISCSLP 2006. This means that the proposed method achieved higher performance than the GMM-based techniques [Zheng 2006; Lee 2006].

Table 2. Evaluation results of the proposed method for CCC Speaker Recognition Evaluation 2006

Identification Correctness Rate	99.33 % (2379/2395)
False Acceptance Rate	0.42 % (10/2395)
False Rejection Rate	0.25 % (6/2395)
Recognition Error Rate	0.00 % (0/2395)

Table 3. Evaluation results using GMM and VQ-distortion for CCC Speaker Recognition Evaluation 2006

Method	GMM	VQ-distortion
Identification Correctness Rate	95.24 % (2281/2395)	96.20 % (2304/2395)
False Acceptance Rate	3.97 % (95/2395)	3.63 % (87/2395)
False Rejection Rate	0.67 % (16/2395)	0.13 % (3/2395)
Recognition Error Rate	0.13 % (3/2395)	0.04 % (1/2395)

For a fair comparison with the proposed method, we conducted experiments using GMM and VQ-distortion based methods using the same feature parameters. Table 3 shows the experimental results. We used diagonal covariance matrices for GMM with 32 mixture components, which obtained the best ICR for testing data with the optimal threshold, *i.e.*, we set the optimal parameters for the GMM and the VQ-distortion based methods posteriorly. The codebook size for the VQ-distortion method was 128.

These results also show the proposed method achieved higher accuracy than the GMM and VQ-distortion methods. Especially, the proposed method reduced the false acceptance of out-of-set speakers.

We expect the reason for these results is the difference between 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. Consequently, by considering the weight of each centroid, the proposed method can avoid the error that occurred with the VQ-distortion based method, *i.e.*, the distortion becomes small because many frames concentrate on one centroid. Due to this, we believe the false acceptance rate of the proposed method was able to be much lower than the conventional methods. On the other hand, the proposed algorithm is computationally expensive. Actually, it took about nine minutes to identify one utterance with 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 was 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 [Fattah 2006A; Park 2002], although it will turn into a language dependent system.

4. Summary

In this paper, we have presented a non-parametric speaker identification method using Earth Mover's Distance (EMD) designed for text-indepedent speaker identification and its evaluation results for *CCC Speaker Recognition Evaluation 2006*, organized by the Chinese Corpus Consortium (CCC) for the *th International Symposium on Chinese Spoken Language Processing* (ISCSLP 2006). The proposed method was originally designed to apply to a distributed speaker recognition system. We have improved the method to be able to handle non-quantized data and reject out-of-set speakers in this paper.

Experimental results, on the text-independent speaker identification task with a closed channel condition, showed the proposed method achieved an identification correctness rate of 99.33 %, which was the best for the task at ISCSLP 2006. This result suggests that the proposed method would also be effective in speaker verification. On the other hand, the proposed method is computationally expensive. We also confirmed that the errors of the proposed method depended on the content of the utterances.

In future work, we will accelerate the distance calculation process in the proposed algorithm and apply the method to speaker verification. Furthermore, we will consider use of speech recognition to improve the speaker identification accuracy. We will also study other distance measures between discrete distributions that are appropriate for speaker recognition.

Acknowledgments

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Exploratory Research, 17656128 and 1930029, 2007, Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B), 17300065, 2007, and International Communications Foundation (ICF). The authors thank Mr. David B. Bracewell for his valuable assistance with English checking. The authors also greatly appreciate the Chinese Corpus Consortium who provided the valuable speech data.

References

Atal, B. S., “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *The Journal of the Acoustical Society of America*, 55(6), 1974, pp. 1304–1312.

Broun, C. C., W. M. Campbell, D. Pearce, and H. Kelleher, “Distributed Speaker Recognition Using the ETSI Distributed Speech Recognition Standard,” *In Proceedings of A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, Crete, Greece, pp. 121–124.

ETSI Standard Document , “Speech processing, transmission and auality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm,” ETSI ES 201 108 v1.1.2, Apr. 2000.

Fattah, M.A., F. Ren, S. Kuroiwa, and I. Fukuda, “Phoneme Based Speaker Modeling to Improve Speaker Recognition,” *Information*, 9(1), 2006A, pp. 135–147.

Fattah, M.A., F. Ren, and S. Kuroiwa, “Effects of Phoneme Type and Frequency on Distributed Speaker Identification and Verification,” *IEICE Transactions on Information and Systems*, E89-D(5), 2006B, pp. 1712–1719.

Fukuda, I., M. A. Fattah, S. Tsuge, and S. Kuroiwa, “Distributed Speaker Identification on Japanese Speech Corpus Using the ETSI Aurora Standard,” *In Proceedings of 3rd International Confference on Information*, 2004, Tokyo, Japan, pp. 207–210.

Grassi, S., M. Ansorge, F. Pellandini, and P.-A. Farine, “Distributed Speaker Recognition Using the ETSI AURORA Standard,” *In Proceedings of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, 2002, Budapest, Hungary, pp. 120–125.

ITU, http://www.itu.int/ITU-T/2001-2004/com16/sg16-q15.htm, 2004.

KDDI News Release, http://www.kddi.com/english/corporate/news_release/2006/0112/, 2006.

Kuroiwa, S., S. Sakayori, S. Yamamoto, and M. Fujioka: "Prank call rejection system for home country direct service," *In Proceedings of IEEE 1st Workshop on Interactive Voice Technology for Telecommunications Applications*, 1996, Basking Ridge, NJ, USA, pp. 135–138.

Kuroiwa, S., Y. Umeda, S. Tsuge, and F. Ren, "Nonparametric Speaker Recognition Method using Earth Mover's Distance," *IEICE Transactions on Information and Systems*, E89-D(3), 2006, pp. 1074–1081.

Lee, K.-A, H. Sun, R. Tong, B. Ma, M. Dong, C. You, D. Zhu, C.-W. Koh, L. Wang, T. Kinnuen, E.-S.Chng, and H. Li, "The IIR Submisson to CSLP 2006 Speaker Recognition Evaluation," *In Proceedings of 5th International Symposiou on Chinese Spoken Language Processing*, LNAI-4274 2006, Singapore, pp. 494–503.

Park, A., and T. Hazen, "ASR dependent techniques for speaker identification," *In Proceedings of 7th International Confference on Spoken Language Processing*, 2002, Denver, CO, USA, pp. 1337–1340.

Pearce, D., "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends," *In Proceedings of Applied Voice Input/Output Society Conference*, 2000, San Jose, CA, USA.

Reynolds, D. A., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, 17(1), 1995, pp. 91–108.

Rubner, Y., L. Guibas, and C. Tomasi, "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval," *In Proceedings of the ARPA Image Understanding Workshop*, 1997, New Orleans, LA, USA, pp. 661–668.

Rubner, Y., http://ai.stanford.edu/ rubner/emd/, 1999.

Sit, C.-H., M.-W. Mak, and S.-Y. Kung, "Maximum Likelihood and Maximum A Posteriori Adaptation for Distributed Speaker Recognition Systems," *In Proceedings of the 1st International Confference on Biometric Authentication*, LNCS-3072 2004, Hong Kong, China, pp. 640–647.

Soong, F., A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* 1985, Tampa, FL, USA, pp. 387–390.

Uchibe, T., S. Kuroiwa, and N. Higuchi, "Determination of threshold for speaker verification using speaker adaptation gain in likelihood during training," *In Proceedings of 6th International Confference on Spoken Language Processing*, 2000, Beijing, China, pp. 326–329.

Young, S. , G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.3), Cambridge University Engineering Depar tment, 2005, http://htk.eng.cam.ac.uk/.

Zheng, T. F., Z. Song, L. Zhang, M. Brasser, W. Wu, and J. Deng, "CCC Speaker Recognition Evaluation 2006: Overview, Methods, Results and Perspective," *In Proceedings of 5th*

International Symposiou on Chinese Spoken Language Processing, LNAI-4274 2006, Singapore, pp. 485–493.