Speaker Recognition using a Non-parametric Speaker Model Representation and Earth Mover's Distance

Yoshiyuki UMEDA[†], Satoru TSUGE[†], Fuji REN[†], and Shingo KUROIWA[†]

† Faculty of Engineering, Tokushima University
2-1 Minami-Josanjima, Tokushima 770-8506, Japan

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 32% for quantized data.

Key words Distributed Speaker Recognition, speaker identification, non-parametric, Earth Mover's Distance

1. Introduction

In recent years, the use of 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 feature. As the 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 security system in this case. 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]~[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

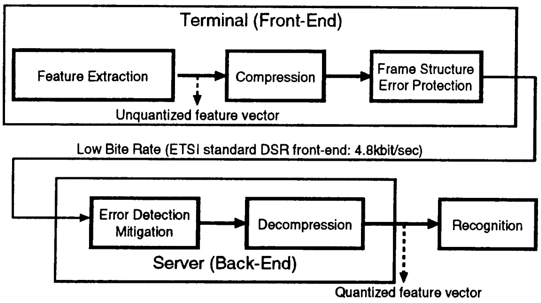

Figure 1 A block diagram of distributed speaker recognition system

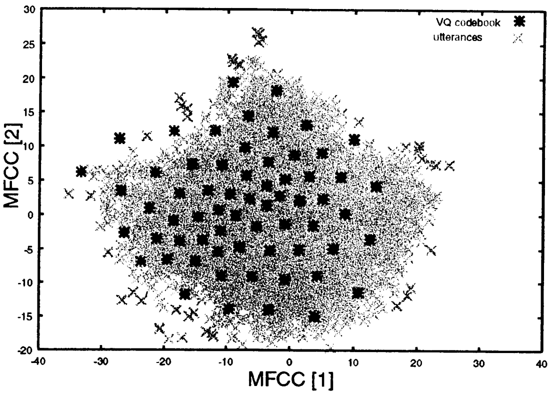

Figure 2 A scatter chart of the feature vectors (1st and 2nd order MFCCs). × is the unquantized feature vector. * is the quantized feature vector.

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(注1) of multidimensional features [7]. Since the proposed method does not need estimation of statistics parameters, it is expected that the proposed method is robust to the quantized data.

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. Section **5.** describes the discussion of these experimental results. Finally, we summarize this paper in section **6.**.

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 and discuss the results of the experiment.

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 [8] 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. Fig. 2 shows the scatter chart of the feature vectors of 25 utterances, which are 1st and 2nd order MFCCs extracted by ETSI standard DSR front-end, and the centroid vectors in codebook which are employed by ETSI DSR front-end. This figure illustrates that many input vectors are quantized into one centroid. In the next section, we investigate the influence of the quantized feature vectors on the 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 section **4. 1.** We used the 16-mixture GMMs in this experiment.

Table 1 shows the experimental results. From this table,

(注1) : In [7], EMD is defined the distance between two *signatures*. Although the *signatures* are histograms that have different bins, we use histogram in this paper

Table 1 The Identification Error Rate (IER) at 8 and 16 kHz

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

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 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 section **2**., we described that it is difficult to use the continuous GMM which is a statistical parametric model as a speaker model for distributed speaker recognition. 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.

First, we provide a brief overview of Earth Mover's Distance. Next, we propose the distributed speaker recognition method using a non-parametric speaker model and EMD measurement and illustrate the flow of this method.

3.1 Earth Mover's Distance

The EMD was proposed by *Rubner et al.* [7] 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 $\boldsymbol{P} = \{(\boldsymbol{p}_1, w_{p_1}), \cdots, (\boldsymbol{p}_m, w_{p_m})\}$ 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 $\boldsymbol{Q} = \{(\boldsymbol{q}_1, w_{q_1}), \cdots, (\boldsymbol{q}_n, w_{q_n})\}$ 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:

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

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

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

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

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:

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

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. The detail of the proposed method is described as follows:

Fig. 3 illustrates the flow of the proposed method.

- 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.

- 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, ten four-digit utterances, and five eight-digit utterances in each session. The average length of the text sentences, the four-digit utterances, and the eight-digit utterances are about five seconds, one seconds, and two seconds, respectively. These utterances were sampled at 16kHz.

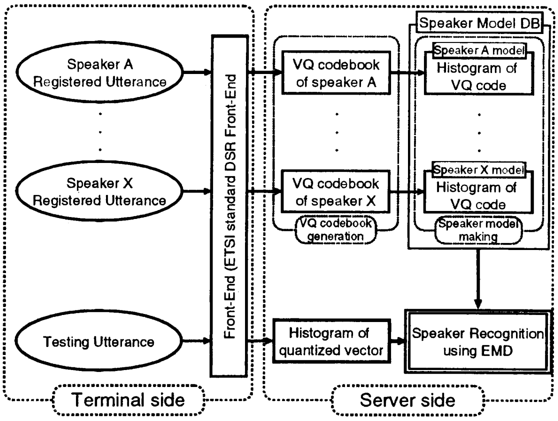

Figure 3 The block diagram of proposed method

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. In this experiment, we defined three test sets. For the first test set, we used five text sentences in six sessions by all speakers, in total 630 utterances (21 speakers × 5 sentences × 6 sessions). The text of these utterances is not contained in the training data. We called this test set "text sentences". For the second test set, we used ten four-digit utterances in six sessions by all speakers, in total 1,260 utterances (21 speakers × 10 sentences × 6 sessions). We called this test set "four-digit". For the third test set, we used five eight-digit utterances in six sessions by all speakers, in total 630 utterances (21 speakers × 5 sentences × 6 sessions). We called this test set "eight-digit". The total number of all testing utterances was 2,520.

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 + 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 the text sentences test set and 1,024 for the four and eight-digit test set. For comparison with the proposed method,

Table 2 The IER for all test set

Method	IER
GMM	11.9%
VQ-distortion	9.0%
EMD-VQ	8.1%

we also conducted experiments with the speaker recognition methods based on GMM and VQ-distortion. 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.

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 and VQ-distortion in the table. The results shown in table 2 are calculated by using all test sets, that is, the text sentences, the four-digit, and eight-digit.

We can see from these results that the VQ-distortion method had a lower IER than the GMM method; 9.0% for VQ-distortion and 11.9% 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. To this end, we conclude that the EMD is a useful distance measure for speaker recognition and the proposed method is an effective method for distributed speaker recognition.

In the following section, we discuss the details of these results.

5. Discussion

This section shows the details of the results of the previous speaker identification experiments and our discussion. We discuss about

- IER for the each test set, text sentences, four-digit, and eight-digit (**5.1**).
- influence of codebook size on IER (**5.2**).

5.1 IER for the each test set

We investigate the influence on the identification performance of each test set. The differences of each test set were the contents of the text and the length of the utterance. The average length of the text sentences was five seconds, the eight-digit was two seconds, and the four-digit was one seconds.

Table 3 The IER for each test set

	test set		
Method	four-digit	eight-digit	text sentences
GMM	16.7%	10.3%	4.0%
VQ distortion	14.8%	5.6%	1.0%
EMD-VQ	13.3%	5.2%	0.6%

Table 3 shows the IER for each test set. First, these results show that the proposed method gave consistently better performance than the conventional methods. Consequently, we conclude again that the proposed method is an effective method for distributed speaker recognition regardless of the length and the phoneme contents of the testing sentence.

From the table, we can also find the difference of IER for each test set. For both the proposed method and conventional methods, the IER decreased as the length of the testing utterances became longer.

Additionally, this table shows that the proposed method gave better improvement in the IER for text sentences than for digit test sets. This result suggests that the proposed method yielded better identification performance if the kind of sentence in the testing data is the same as in the registered data, i.e., the phoneme occurrence frequency is similar between the testing data and the registered data. Comparing with the VQ distortion method, the proposed method makes the best use of the distribution of the speaker model, which is the occurence frequency of each cluster (w_i in section **3.1**) for speaker identification. Hence, we are planning to study the relationship between the phoneme occurrence frequency and weight, w_i, in the proposed method.

5.2 Influence of VQ codebook size on IER

In the above experiments, we used the optimum codebook size which indicates the best performance for each test set. We investigated the influence of the VQ codebook size on the identification performance. Fig. 4 and Fig. 5 show the IER of the proposed method for each test set as a function of the number of VQ centroids. From these figures, we can see that the optimum codebook size changes with each test

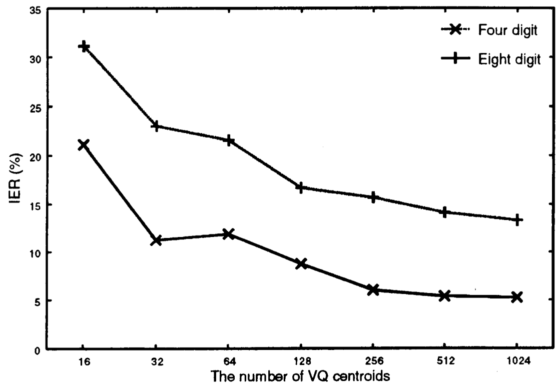

Figure 4 IER for digit test set as a function of the number of VQ centroids

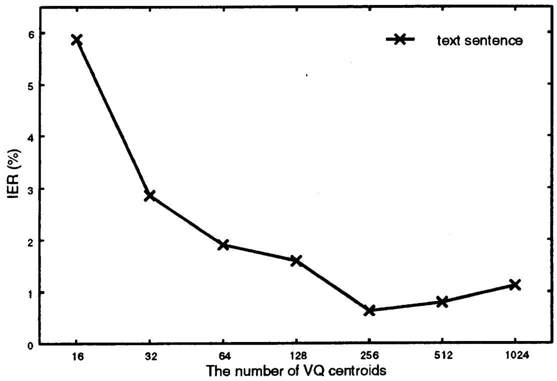

Figure 5 IER for text sentences test set as a function of the number of VQ centroids

set. The optimum codebook size (the number of centroids) is 256 under the condition of text sentences test set, while it is 1,024 under the condition of digit test sets. This tendency was also observed in the VQ-distortion method. When a test utterance and registered utterance are similar, the VQ codebook of small size may be sufficient. On the other hand, the proposed method and VQ-distortion method may require a large VQ codebook for different kinds of utterances. Now, we are conducting further investigations to study the decision function of optimum VQ codebook size.

6. 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 (31.9% over the GMM method and 9.1% over the VQ-distortion method) by using the proposed method. We also confirmed that the proposed method is more effective when the kind of registered utterance is the same as testing data.

In this paper, we used five text utterances for the registered data. We are now planning to conduct further experiments by using digit utterance for the registered data. In future work, we will evaluate the performance of the proposed method using a larger database, and apply the proposed method to speaker verification.

7. 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, Hoso-Bunka Foundation (HBF) and International Communications Foundation (ICF).

References

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

[2] "ETSI ES 201 108 v1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithm," 2000.

[3] I. Fukuda, S. Tsuge, S. Kuroiwa, and F. Ren, "A speaker recognitino using ETSI standard DSR front-end," J. Acoust. Soc. Jap., pp.175–176, 2004. (in Japanese).

[4] C. Broun, W. 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, 2001.

[5] C. Sit, M. Mak, and S. Kung, "Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems," ICBA, 2004.

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

[7] Y. Rubner, L. Guibas, and C. Tomasi, "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval," Proc. of the ARPA Image Understanding Workshop, pp.661–668, 1997.

[8] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," ISCA ITRW ASR, pp.191–188, 2000.