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.

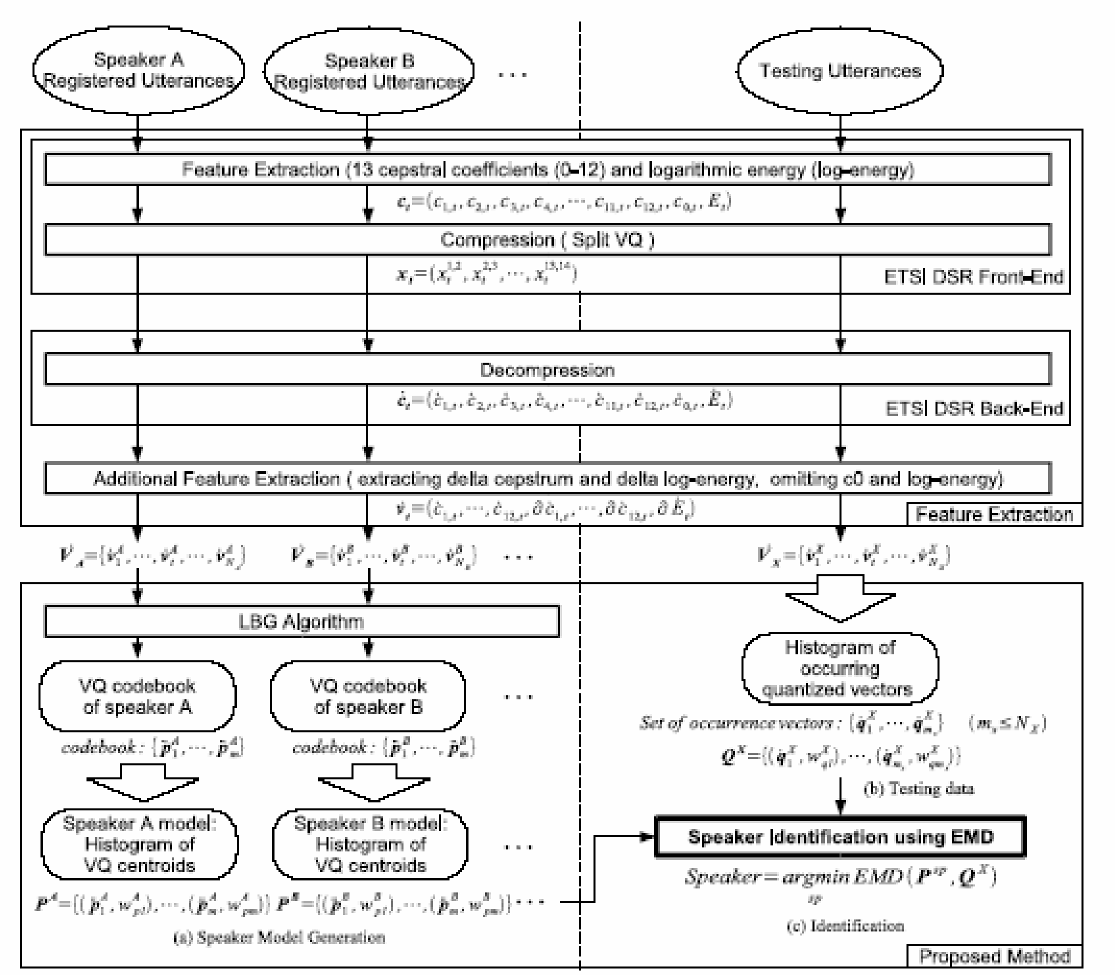

Speaker A Registered Utterances
Speaker B Registered Utterances
. . .
Testing Utterances
Feature Extraction (13 cepstral coefficients (0-12) and logarithmic energy (log-energy)
$c_t=(c_{1,t}, c_{2,t}, c_{3,t}, c_{4,t}, \cdots, c_{11,t}, c_{12,t}, c_{0,t}, E_t)$
Compression (Split VQ)
$x_t=(x_t^{1,2}, x_t^{2,3}, \cdots, x_t^{13,14})$
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 : $\{\dot{q}_1^X, \cdots, \dot{q}_{m_x}^X\}$ $(m_x \leq N_X)$
$Q^X=\{(\dot{q}_1^X, w_{q1}^X), \cdots, (\dot{q}_{m_x}^X, w_{qm_x}^X)\}$
(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, 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); $\dot{c}_t$ is a decompressed feature vector; and $\dot{v}_t$ 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.

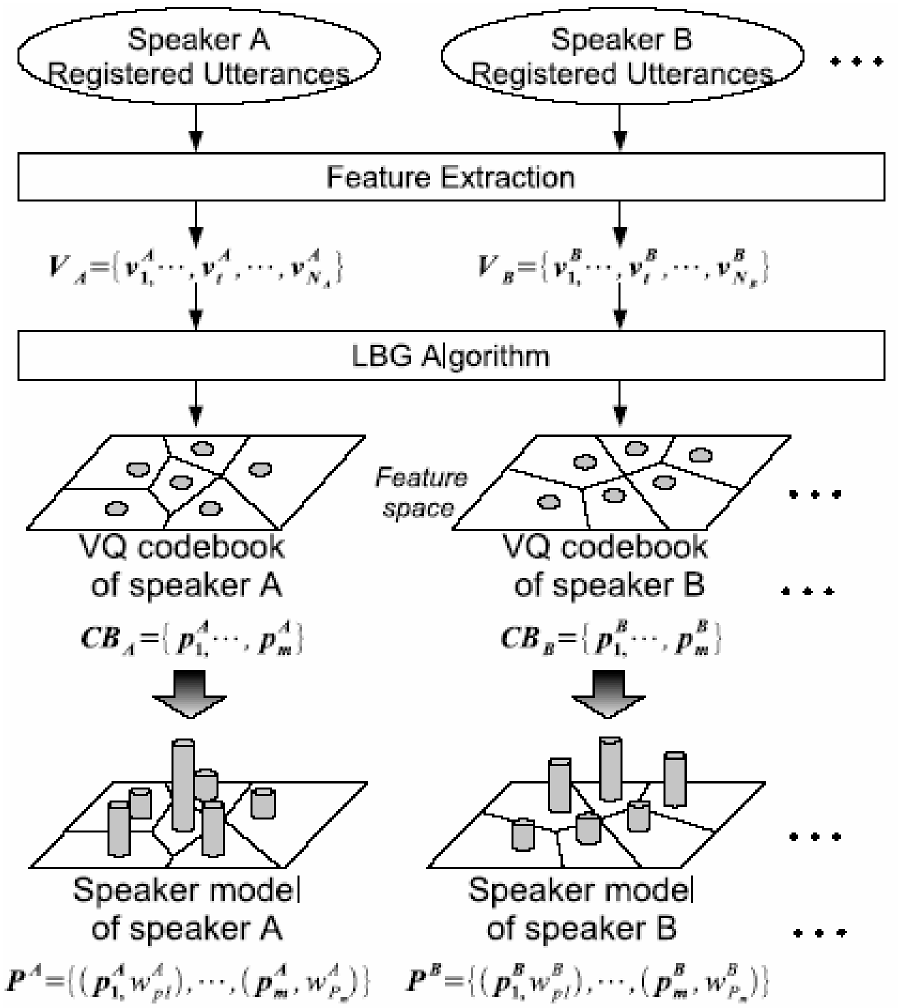
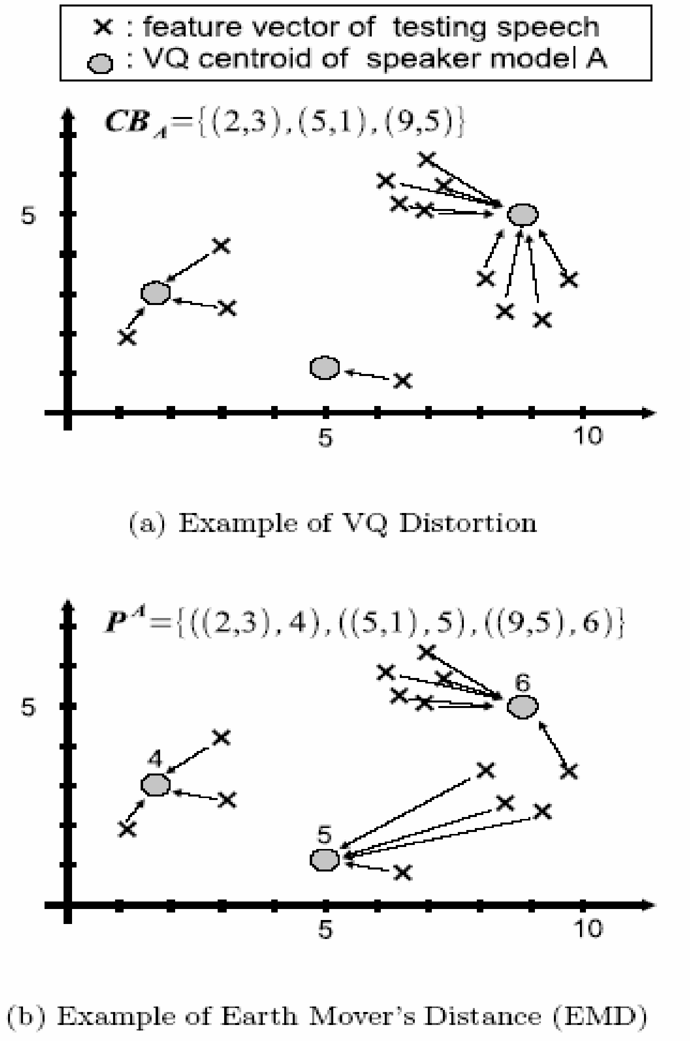

Speaker A Registered Utterances
Speaker B Registered Utterances
Feature Extraction
$V_A=\{v_1^A,\cdots,v_t^A,\cdots,v_{N_A}^A\}$
$V_B=\{v_1^B,\cdots,v_t^B,\cdots,v_{N_B}^B\}$
LBG Algorithm
Feature space
VQ codebook of speaker A
VQ codebook of speaker B
$CB_A=\{p_1^A,\cdots,p_m^A\}$
$CB_B=\{p_1^B,\cdots,p_m^B\}$
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
P^A={((2,3),4),((5,1),5),((9,5),6)}
6
4
5
5
5
10

(a) Example of VQ Distortion

(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.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 255-272

A Novel Characterization of the Alternative Hypothesis Using Kernel Discriminant Analysis for LLR-Based Speaker Verification

Yi-Hsiang Chao*+ , Hsin-Min Wang* and Ruei-Chuan Chang*+

Abstract

In a log-likelihood ratio (LLR)-based speaker verification system, the alternative hypothesis is usually difficult to characterize a priori, since the model should cover the space of all possible impostors. In this paper, we propose a new LLR measure in an attempt to characterize the alternative hypothesis in a more effective and robust way than conventional methods. This LLR measure can be further formulated as a non-linear discriminant classifier and solved by kernel-based techniques, such as the Kernel Fisher Discriminant (KFD) and Support Vector Machine (SVM). The results of experiments on two speaker verification tasks show that the proposed methods outperform classical LLR-based approaches.

Keywords: Kernel Fisher Discriminant, Log-likelihood Ratio, Speaker Verification, Support Vector Machine.

1. Introduction

In essence, the speaker verification task is a hypothesis testing problem. Given an input utterance *U*, the goal is to determine whether *U* was spoken by the hypothesized speaker or not. The log-likelihood ratio (LLR)-based detector [Reynolds 1995] is one of the state-of-the-art approaches for speaker verification. Consider the following hypotheses:

H_0: *U* is from the hypothesized speaker,
H_1: *U* is not from the hypothesized speaker.

The LLR test is expressed as:

* Institute of Information Science, Academia Sinica, Taipei, Taiwan
+ Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
E-mail: {yschao,whm}@iis.sinica.edu.tw; rc@cc.nctu.edu.tw

[Received April 3, 2007; Revised August 9, 2007; Accepted August 12, 2007]

$$L(U) = \log \frac{p(U \mid H_0)}{p(U \mid H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{accept } H_1 \text{ (i.e., reject } H_0 \text{)} , \end{cases} \tag{1}$$

where $p(U \mid H_i)$, $i = 0, 1$, is the likelihood of hypothesis H_i given the utterance U, and θ is the threshold. H_0 and H_1 are, respectively, called the null hypothesis and the alternative hypothesis. Mathematically, H_0 and H_1 can be represented by parametric models denoted as λ and $\bar{\lambda}$, respectively; $\bar{\lambda}$ is often called an anti-model. Though H_0 can be modeled straightforwardly using speech utterances from the hypothesized speaker, H_1 does not involve any specific speaker, thus lacks explicit data for modeling. Many approaches have been proposed to characterize H_1, and various LLR measures have been developed. We can formulate these measures in the following general form [Reynolds 2000]:

$$L(U) = \log \frac{p(U \mid \lambda)}{p(U \mid \bar{\lambda})} = \log \frac{p(U \mid \lambda)}{\Psi(p(U \mid \lambda_1), p(U \mid \lambda_2), \ldots, p(U \mid \lambda_N))}, \tag{2}$$

where $\Psi(\cdot)$ is some function of the likelihood values from a set of so-called background models $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. For example, the background model set can be obtained from N representative speakers, called a cohort [Rosenberg 1992], which simulates potential impostors. If $\Psi(\cdot)$ is an average function [Reynolds 1995], the LLR can be written as:

$$L_1(U) = \log p(U \mid \lambda) - \log \left\{ \frac{1}{N} \sum_{i=1}^{N} p(U \mid \lambda_i) \right\}. \tag{3}$$

Alternatively, the average function can be replaced by various functions, such as the maximum [Liu 1996], *i.e.*:

$$L_2(U) = \log p(U \mid \lambda) - \max_{1 \leq i \leq N} \log p(U \mid \lambda_i), \tag{4}$$

or the geometric mean [Liu 1996], *i.e.*,

$$L_3(U) = \log p(U \mid \lambda) - \frac{1}{N} \sum_{i=1}^{N} \log p(U \mid \lambda_i). \tag{5}$$

A special case arises when $\Psi(\cdot)$ is an identity function and $N = 1$. In this instance, a single background model is usually trained by pooling all the available data, which is generally irrelevant to the clients, from a large number of speakers. This is called the world model or the Universal Background Model (UBM) [Reynolds 2000]. The LLR in this case becomes:

$$L_4(U) = \log p(U \mid \lambda) - \log p(U \mid \Omega), \tag{6}$$

where Ω denotes the world model.

However, none of the LLR measures developed so far has proven to be absolutely superior to any other, since the selection of $\Psi(\cdot)$ is usually application and training data dependent. In particular, the use of a simple function, such as the average, maximum, or geometric mean, is a heuristic that does not include any optimization process. The issues of selection, size, and combination of background models motivate us to design a more comprehensive function, $\Psi(\cdot)$, to improve the characterization of the alternative hypothesis. In this paper, we first propose a new LLR measure in an attempt to characterize H_1 by integrating all the background models in a more effective and robust way than conventional methods. Then, we formulate this new LLR measure as a non-linear discriminant classifier and apply kernel-based techniques, including the Kernel Fisher Discriminant (KFD) [Mika 1999] and Support Vector Machine (SVM) [Burges 1998], to optimally separate the LLR samples of the null hypothesis from those of the alternative hypothesis.

SVM-based techniques have been successfully applied to many classification and regression tasks, including speaker verification. Unlike our work, existing approaches [Bengio 2001; Wan 2005] only use a single background model, *i.e.*, the world model, to represent the alternative hypothesis, instead of integrating multiple background models to characterize the alternative hypothesis. For example, Bengio *et al.* [Bengio 2001] proposed a decision function:

$$L_5(U) = a_1 \log p(U \mid \lambda) - a_2 \log p(U \mid \Omega) + b, \tag{7}$$

where a_1, a_2, and b are adjustable parameters estimated using SVM. An extended version of Eq. (7) with the Fisher kernel and the LR score-space kernel for SVM was investigated in Wan [Wan 2005].

The results of speaker verification experiments conducted on both the XM2VTS database [Messer 1999] and the ISCSLP2006-SRE database [Chinese Corpus Consortium 2006] show that the proposed methods outperform classical LLR-based approaches. The remainder of this paper is organized as follows. Section 2 describes the design of the new LLR measure in our approach. Sections 3 and 4 introduce the kernel discriminant analysis used in this work and the formation of the characteristic vector by background model selection, respectively. Section 5 contains our experiment results. Finally, in Section 6, we present our conclusions.

2. New LLR Measure Design

2.1 Analysis of the Alternative Hypothesis

First of all, we redesign the function $\Psi(\cdot)$ in Eq. (2) as:

$$p(U \mid \bar{\lambda}) = \Psi(\mathbf{u}) = (p(U \mid \lambda_1)^{w_1} \cdot p(U \mid \lambda_2)^{w_2} \cdot \ldots \cdot p(U \mid \lambda_N)^{w_N})^{1/(w_1+w_2+\ldots+w_N)}, \tag{8}$$

where $\mathbf{u} = [p(U \mid \lambda_1), p(U \mid \lambda_2), ..., p(U \mid \lambda_N)]^T$ is an $N \times 1$ vector and w_i is the weight of the likelihood $p(U \mid \lambda_i)$, i = 1,2,..., N. This function gives N background models different weights according to their individual contribution to the alternative hypothesis. It is clear that Eq. (8) is equivalent to a geometric mean function when $w_i = 1$, i = 1,2,..., N. If some background model λ_i contrasts with an input utterance U, the likelihood $p(U \mid \lambda_i)$ may be extremely small, thus causing the geometric mean to approximate zero. In contrast, by assigning a favorable weight to each background model, the function $\Psi(\cdot)$ defined in Eq. (8) may be less affected by any specific background model with an extremely small likelihood. Therefore, the resulting score for the alternative hypothesis obtained by Eq. (8) will be more robust and reliable than that obtained by a geometric mean function. It is also clear that Eq. (8) will reduce to a maximum function when $w_{i*} = 1$, $i^* = \arg\max_{1 \le i \le N} \log p(U \mid \lambda_i)$; and $w_i = 0$, $\forall i \ne i^*$.

By substituting Eq. (8) into Eq. (2), we obtain:

$$\begin{aligned} L_6(U) &= \log \frac{p(U \mid \lambda)}{p(U \mid \bar{\lambda})} \\ &= \log \frac{p(U \mid \lambda)}{(p(U \mid \lambda_1)^{w_1} \cdot p(U \mid \lambda_2)^{w_2} \cdot ... \cdot p(U \mid \lambda_N)^{w_N})^{1/(w_1+w_2+...+w_N)}} \\ &= \log \left(\left(\frac{p(U \mid \lambda)}{p(U \mid \lambda_1)} \right)^{w_1} \cdot \left(\frac{p(U \mid \lambda)}{p(U \mid \lambda_2)} \right)^{w_2} \cdot ... \cdot \left(\frac{p(U \mid \lambda)}{p(U \mid \lambda_N)} \right)^{w_N} \right)^{1/(w_1+w_2+...+w_N)} \\ &= \frac{1}{w_1 + w_2 + ... + w_N} \left(w_1 \log \frac{p(U \mid \lambda)}{p(U \mid \lambda_1)} + w_2 \log \frac{p(U \mid \lambda)}{p(U \mid \lambda_2)} + ... + w_N \log \frac{p(U \mid \lambda)}{p(U \mid \lambda_N)} \right) \\ &= \frac{1}{w_1 + w_2 + ... + w_N} \mathbf{w}^T \mathbf{x} \begin{cases} \ge \theta & \text{accept} \\ < \theta & \text{reject} \end{cases} \\ &= \mathbf{w}^T \mathbf{x} \begin{cases} \ge \theta' & \text{accept} \\ < \theta' & \text{reject,} \end{cases} \end{aligned} \tag{9}$$

where $\mathbf{w} = [w_1, w_2 ..., w_N]^T$ is an $N \times 1$ weight vector, the new threshold $\theta' = (w_1 + w_2 + ... + w_N)\theta$, and $\mathbf{x}$ is an $N \times 1$ vector in the space R^N, expressed by

$$\mathbf{x} = [\log \frac{p(U \mid \lambda)}{p(U \mid \lambda_1)}, \log \frac{p(U \mid \lambda)}{p(U \mid \lambda_2)}, ..., \log \frac{p(U \mid \lambda)}{p(U \mid \lambda_N)}]^T. \tag{10}$$

The implicit idea in Eq. (10) is that the speech utterance U can be represented by a characteristic vector $\mathbf{x}$.

If we replace the threshold θ' in Eq. (9) with a bias b, the equation can be rewritten as:

$$L(U) = \mathbf{w}^T \mathbf{x} + b = f(\mathbf{x}), \tag{11}$$

where *f*(**x**) forms a so-called linear discriminant classifier. This classifier translates the goal of solving an LLR measure into the optimization of **w** and *b*, such that the utterances of clients and impostors can be separated. To realize this classifier, three distinct data sets are needed: one for generating each client's model, one for generating the background models, and one for optimizing **w** and *b*. Since the bias *b* plays the same role as the decision threshold of the conventional LLR measure, which can be determined through a trade-off between false acceptance and false rejection, the main goal here is to find **w**. Existing linear discriminant analysis techniques, such as Fisher's Linear Discriminant (FLD) [Duda 2001] or Linear SVM [Burges 1998], can be applied to implement Eq. (11).

2.2 Linear Discriminant Analysis

Fisher's Linear Discriminant (FLD) is one of the popular linear discriminant classifiers [Duda 2001]. Suppose the *i*-th class has n_i data samples, $\mathbf{X}_i = \{\mathbf{x}_1^i,..,\mathbf{x}_{n_i}^i\}$, *i* = 1, 2. The goal of FLD is to seek a direction **w** in the space R^N such that the following Fisher's criterion function *J*(**w**) is maximized:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}, \tag{12}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are, respectively, the between-class scatter matrix and the within-class scatter matrix defined as

$$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \tag{13}$$

and

$$\mathbf{S}_w = \sum_{i=1,2} \sum_{\mathbf{x}\in\mathbf{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \tag{14}$$

where $\mathbf{m}_i$ is the mean vector of the *i*-th class computed by

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{s=1}^{n_i} \mathbf{x}_s^i. \tag{15}$$

According to Duda [Duda 2001], the solution for **w**, which maximizes *J*(**w**) defined in Eq. (12), is the leading eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

3. Kernel Discriminant Analysis

Intuitively, *f*(**x**) in Eq. (11) can be solved via linear discriminant training algorithms [Duda 2001], such as FLD or Linear SVM. However, such methods are based on the assumption that the observed data of different classes is linearly separable, which is obviously not feasible in most practical cases with nonlinearly separable data. To solve this problem more effectively, we propose using a kernel-based nonlinear discriminant classifier. It is hoped that data from different classes, which is not linearly separable in the original input space R^N, can be separated linearly in a certain higher dimensional (maybe infinite) feature space F via a nonlinear mapping Φ. Let Φ(**x**) denote a vector obtained by mapping **x** from R^N to F. Then, the objective function, based on Eq. (11), can be re-defined as:

$$f(\mathbf{x}) = \mathbf{w}_F{}^T \Phi(\mathbf{x}) + b\ , \tag{16}$$

which constitutes a linear discriminant classifier in F, where $\mathbf{w}_F$ is a weight vector in F.

In practice, it is difficult to determine the kind of mapping that would be applicable; therefore, the computation of Φ(**x**) might be infeasible. To overcome this difficulty, a promising approach is to characterize the relationship between the data samples in F, instead of computing Φ(**x**) directly. This is achieved by introducing a kernel function k(**x**, **y**)=<Φ(**x**),Φ(**y**)>, which is the dot product of two vectors Φ(**x**) and Φ(**y**) in F. The kernel function $k(\cdot)$ must be symmetric, positive definite and conform to Mercer's condition [Burges 1998].

A number of kernel functions exist, such as the simplest dot product kernel function k(**x**, **y**) = $\mathbf{x}^T\mathbf{y}$, and the very popular Radial Basis Function (RBF) kernel k(**x**, **y**) = exp(− $\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2$) in which σ is a tunable parameter. Existing techniques, such as KFD [Mika 1999] or SVM [Burges 1998], can be applied to implement Eq. (16).

3.1 Kernel Fisher Discriminant (KFD)

Suppose the i-th class has n_i data samples, $\mathbf{X}_i = \{\mathbf{x}_1^i,..,\mathbf{x}_{n_i}^i\}$, i = 1, 2. The goal of KFD is to seek a direction $\mathbf{w}_F$ in the feature space F such that the following Fisher's criterion function $J(\mathbf{w}_F)$ is maximized:

$$J(\mathbf{w}_F) = \frac{\mathbf{w}_F{}^T \mathbf{S}_b^\Phi \mathbf{w}_F}{\mathbf{w}_F{}^T \mathbf{S}_w^\Phi \mathbf{w}_F}, \tag{17}$$

where $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ are, respectively, the between-class scatter matrix and the within-class scatter matrix in F defined as:

$$\mathbf{S}_b^\Phi = (\mathbf{m}_1^\Phi - \mathbf{m}_2^\Phi)(\mathbf{m}_1^\Phi - \mathbf{m}_2^\Phi)^T \tag{18}$$

and

$$\mathbf{S}_w^\Phi = \sum_{i=1,2} \sum_{\mathbf{x}\in\mathbf{X}_i} (\Phi(\mathbf{x}) - \mathbf{m}_i^\Phi)(\Phi(\mathbf{x}) - \mathbf{m}_i^\Phi)^T, \tag{19}$$

where $\mathbf{m}_i^\Phi = (1/n_i)\sum_{s=1}^{n_i} \Phi(\mathbf{x}_s^i)$, and i = 1, 2, is the mean vector of the i-th class in F. Let $\mathbf{X}_1 \cup \mathbf{X}_2 = \{\mathbf{x}_1^1,..,\mathbf{x}_{n_1}^1\} \cup \{\mathbf{x}_1^2,..,\mathbf{x}_{n_2}^2\} = \{\mathbf{x}_1,..,\mathbf{x}_l\}$ and $l = n_1 + n_2$. Since the solution of $\mathbf{w}_F$ must lie in the span of all training data samples mapped in F [Mika 1999], $\mathbf{w}_F$ can be expressed as:

$$\mathbf{w}_F = \sum_{j=1}^{l} \alpha_j \Phi(\mathbf{x}_j). \tag{20}$$

Let $\boldsymbol{\alpha}^T = [\alpha_1, \alpha_2,..., \alpha_l]$. Accordingly, Eq. (16) can be re-written as:

$$f(\mathbf{x}) = \sum_{j=1}^{l} \alpha_j k(\mathbf{x}_j, \mathbf{x}) + b. \tag{21}$$

Our goal, therefore, changes from finding $\mathbf{w}_F$ to finding $\boldsymbol{\alpha}$, which maximizes

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{N} \boldsymbol{\alpha}}, \tag{22}$$

where $\mathbf{M}$ and $\mathbf{N}$ are computed by:

$$\mathbf{M} = (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)^T \tag{23}$$

and

$$\mathbf{N} = \sum_{i=1,2} \mathbf{K}_i (\mathbf{I}_{n_i} - \mathbf{1}_{n_i}) \mathbf{K}_i^T, \tag{24}$$

respectively, where $\boldsymbol{\eta}_i$ is an l×1 vector whose j-th element $(\boldsymbol{\eta}_i)_j = (1/n_i)\sum_{s=1}^{n_i} k(\mathbf{x}_j, \mathbf{x}_s^i)$, j = 1,2,..., l; $\mathbf{K}_i$ is an $l \times n_i$ matrix with $(\mathbf{K}_i)_{js} = k(\mathbf{x}_j, \mathbf{x}_s^i)$; $\mathbf{I}_{n_i}$ is an $n_i \times n_i$ identity matrix; and $\mathbf{1}_{n_i}$ is an $n_i \times n_i$ matrix with all entries equal to $1/n_i$. Following Mika [Mika 1999], the solution for $\boldsymbol{\alpha}$, which maximizes $J(\boldsymbol{\alpha})$ defined in Eq. (22), is the leading eigenvector of $\mathbf{N}^{-1}\mathbf{M}$.

3.2 Support Vector Machine (SVM)

Alternatively, Eq. (16) can be solved with an SVM, the goal of which is to seek a separating hyperplane in the feature space F that maximizes the margin between classes. Following Burges [Burges 1998], $\mathbf{w}_F$ is expressed as:

$$\mathbf{w}_F = \sum_{j=1}^{l} y_j \alpha_j \Phi(\mathbf{x}_j), \tag{25}$$

which yields

$$f(\mathbf{x}) = \sum_{j=1}^{l} y_j \alpha_j k(\mathbf{x}_j, \mathbf{x}) + b, \tag{26}$$

where each training sample $\mathbf{x}_j$ belongs to one of the two classes identified by the label $y_j \in \{-1,1\}$, j=1, 2,..., l. We can find the coefficients α_j by maximizing the objective function,

$$Q(\boldsymbol{\alpha}) = \sum_{j=1}^{l} \alpha_j - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j), \tag{27}$$

subject to the constraints,

$$\sum_{j=1}^{l} y_j \alpha_j = 0, \text{ and } 0 \le \alpha_j \le C_\alpha, \ \forall j, \tag{28}$$

where C_α is a penalty parameter [Burges 1998]. The problem can be solved using quadratic programming techniques [Vapnik 1998]. Note that most α_j are equal to zero, and the training samples associated with non-zero α_j are called *support vectors*. A few support vectors act as the key to deciding the optimal margin between classes in the SVM. An SVM with a dot product kernel function is known as a Linear SVM.

4. Formation of the Characteristic Vector

In our experiments, we use B+1 background models, consisting of B cohort set models and one world model, to form the characteristic vector **x** in Eq. (10); and B cohort set models for $L_1(U)$ in Eq. (3), $L_2(U)$ in Eq. (4), and $L_3(U)$ in Eq. (5). Two cohort selection methods [Reynolds 1995] are used in the experiments. One selects the B closest speakers to each client; and the other selects the B/2 closest speakers to, plus the B/2 farthest speakers from, each client. The selection is based on the speaker distance measure [Reynolds 1995], computed by:

$$d(\lambda_i, \lambda_j) = \log\frac{p(U_i \mid \lambda_i)}{p(U_i \mid \lambda_j)} + \log\frac{p(U_j \mid \lambda_j)}{p(U_j \mid \lambda_i)}, \tag{29}$$

where λ_i and λ_j are speaker models trained using the i-th speaker's utterances U_i and the j-th speaker's utterances U_j, respectively. Two cohort selection methods yield the following two (B+1)×1 characteristic vectors:

$$\mathbf{x} = \left[\log\frac{p(U \mid \lambda)}{p(U \mid \Omega)} \ \ \log\frac{p(U \mid \lambda)}{p(U \mid \lambda_{\text{cst }1})} \ \ldots \ \log\frac{p(U \mid \lambda)}{p(U \mid \lambda_{\text{cst }B})} \right]^T \tag{30}$$

and

$$\mathbf{x} = [\log\frac{p(U|\lambda)}{p(U|\Omega)} \log\frac{p(U|\lambda)}{p(U|\lambda_{\text{cst}1})} \cdots \log\frac{p(U|\lambda)}{p(U|\lambda_{\text{cst}B/2})} \log\frac{p(U|\lambda)}{p(U|\lambda_{\text{fst}1})} \cdots \log\frac{p(U|\lambda)}{p(U|\lambda_{\text{fst}B/2})}]^T , \tag{31}$$

where $\lambda_{\text{cst } i}$ and $\lambda_{\text{fst } i}$ are the i-th closest model and the i-th farthest model of the client model λ, respectively.

5. Experiments

We evaluate the proposed approaches on two databases: the XM2VTS database [Messer 1999] and the ISCSLP2006 speaker recognition evaluation (ISCSLP2006-SRE) database [Chinese Corpus Consortium 2006].

For the performance evaluation, we adopt the Detection Error Tradeoff (DET) curve [Martin 1997]. In addition, the NIST Detection Cost Function (DCF), which reflects the performance at a single operating point on the DET curve, is also used. The DCF is defined as:

$$C_{DET} = C_{Miss} \times P_{Miss} \times P_{T\arg et} + C_{FalseAlarm} \times P_{FalseAlarm} \times (1 - P_{T\arg et}), \tag{32}$$

where P_{Miss} and $P_{FalseAlarm}$ are the miss probability and the false-alarm probability, respectively, C_{Miss} and $C_{FalseAlarm}$ are the respective relative costs of detection errors, and $P_{T\arg et}$ is the *a priori* probability of the specific target speaker. A special case of the DCF is known as the Half Total Error Rate (HTER), where C_{Miss} and $C_{FalseAlarm}$ are both equal to 1, and $P_{T\arg et} = 0.5$, i.e., $\text{HTER} = (P_{Miss} + P_{FalseAlarm})/2$.

5.1 Evaluation on the XM2VTS Database

The first set of speaker verification experiments was conducted on speech data extracted from the XM2VTS database [Messer 1999], which is a multimodal database consisting of face images, video sequences, and speech recordings taken on 295 subjects. The raw database contained approximately 30 hours of digital video recordings, which was then manually annotated. Each subject participated in four recording sessions at approximately one-month intervals, and each recording session consisted of two shots. In a shot, every subject was prompted to read three sentences “0 1 2 3 4 5 6 7 8 9”, “5 0 6 9 2 8 1 3 7 4”, and “Joe took father’s green shoe bench out” at his/her normal pace. The speech was recorded by a microphone clipped to the subject’s shirt.

In accordance with Configuration II of the evaluation protocol described in Luettin [Luettin 1998], the XM2VTS database was divided into three subsets: “Training”, “Evaluation”, and “Test”. In our speaker verification experiments, we used the “Training” subset to build the individual client’s model and the world model[1], and the “Evaluation” subset to estimate the decision threshold θ in Eq. (1) and the parameters $\mathbf{w}$, $\mathbf{w}_F$, and b in

[1] Currently, we do not have an external resource to train the world model and the background models. We follow the evaluation protocol in [Luettin 1998], which suggests “If a world model is needed, as in speaker verification, a client-dependent world model can be trained from all other clients but the actual client. Although not optimal, it is a valid method.” We will train the world model and the background models using an external resource in our future work.

Eq. (11) or Eq. (16). The performance of speaker verification was then evaluated on the "Test" subset. As shown in Table 1, a total of 293 speakers[2] in the database were divided into 199 clients, 25 "evaluation impostors", and 69 "test impostors".

Table 1. Configuration II of the XM2VTS database.

Session	Shot	199 clients	25 evaluation impostors	69 test impostors
1	1	Training	Evaluation	Test
	2			
2	1			
	2			
3	1	Evaluation		
	2			
4	1	Test		
	2			

We used 12 (2×2×3) utterances/speaker from sessions 1 and 2 to train the individual client's model, represented by a Gaussian Mixture Model (GMM) [Reynolds 1995] with 64 mixture components. For each client, the other 198 clients' utterances from sessions 1 and 2 were used to generate the world model, represented by a GMM with 256 mixture components; 20 or 40 speakers were chosen from these 198 clients as the cohort. Then, we used 6 utterances/client from session 3, and 24 (4×2×3) utterances/evaluation-impostor over the four sessions, which yielded 1,194 (6×199) client samples and 119,400 (24×25×199) impostor samples, to estimate θ, $\mathbf{w}$, $\mathbf{w}_F$, and b. However, as a kernel-based classifier can be intractable when a large number of training samples is involved, we reduced the number of impostor samples from 119,400 to 2,250 using a uniform random selection method. In the performance evaluation, we tested 6 utterances/client in session 4 and 24 utterances/test-impostor over the four sessions, which produced 1,194 (6×199) client trials and 329,544 (24×69×199) impostor trials. Table 2 summarizes all the parametric models used in each system.

Using a 32-ms Hamming-windowed frame with 10-ms shifts, each speech utterance (sampled at 32 kHz) was converted into a stream of 24-order feature vectors, each consisting of 12 Mel-scale frequency cepstral coefficients [Huang 2001] and their first time derivatives.

[2] We discarded 2 speakers (ID numbers 313 and 342) because of partial data corruption.

Table 2. A summary of the parametric models used in each system for the XM2VTS task.

System	H_0	H_1	
	a 64-mixture client GMM	a 256-mixture world model	B 64-mixture cohort GMMs
L_1	√		√
L_2	√		√
L_3	√		√
L_4	√	√	
L_5	√	√	
L_6	√	√	√

5.1.1 Experiment Results

First, B was set to 20 in the experiments. We implemented the proposed LLR system based on linear-based classifiers (FLD and Linear SVM) and kernel-based classifiers (KFD and SVM) in eight ways: 1) FLD with Eq. (30) ("FLD_w_20c"), 2) FLD with Eq. (31) ("FLD_w_10c_10f"), 3) Linear SVM with Eq. (30) ("LSVM_w_20c"), 4) Linear SVM with Eq. (31) ("LSVM_w_10c_10f"), 5) KFD with Eq. (30) ("KFD_w_20c"), 6) KFD with Eq. (31) ("KFD_w_10c_10f"), 7) SVM with Eq. (30) ("SVM_w_20c"), and 8) SVM with Eq. (31) ("SVM_w_10c_10f"). Both SVM and KFD used an RBF kernel function with σ= 5. For performance comparison, we used six systems as our baselines: 1) $L_1(U)$ with the 20 closest cohort models ("*L*1_20c"), 2) $L_1(U)$ with the 10 closest cohort models plus the 10 farthest cohort models ("*L*1_10c_10f"), 3) $L_2(U)$ with the 20 closest cohort models ("*L*2_20c"), 4) $L_3(U)$ with the 20 closest cohort models ("*L*3_20c"), 5) $L_4(U)$ ("*L*4"), and 6) $L_5(U)$ using an RBF kernel function with σ= 10 ("*L*5").

Figure 1 shows the results of the baseline systems evaluated on the "Test" subset in DET curves. We observe that the curves "*L*1_10c_10f", "*L*4" and "*L*5" are better than the others. Thus, in the subsequent experiments, we focused on the performance improvements of our proposed LLR systems over these three baselines.

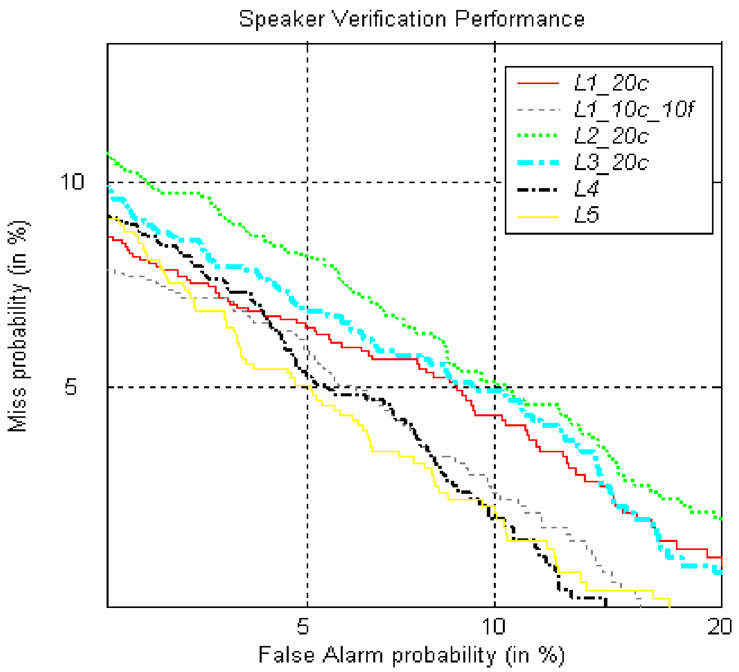


Figure 1. Baselines: DET curves for the XM2VTS "Test" subset (B = 20).

The results of our proposed LLR systems, based on linear-based classifiers and kernel-based classifiers, versus the baseline systems evaluated on the "Test" subset are shown in Figs. 2 and 3, respectively. It is clear that the proposed LLR systems based on either linear-based classifiers or kernel-based classifiers outperform the baseline systems, while KFD perform better than SVM.

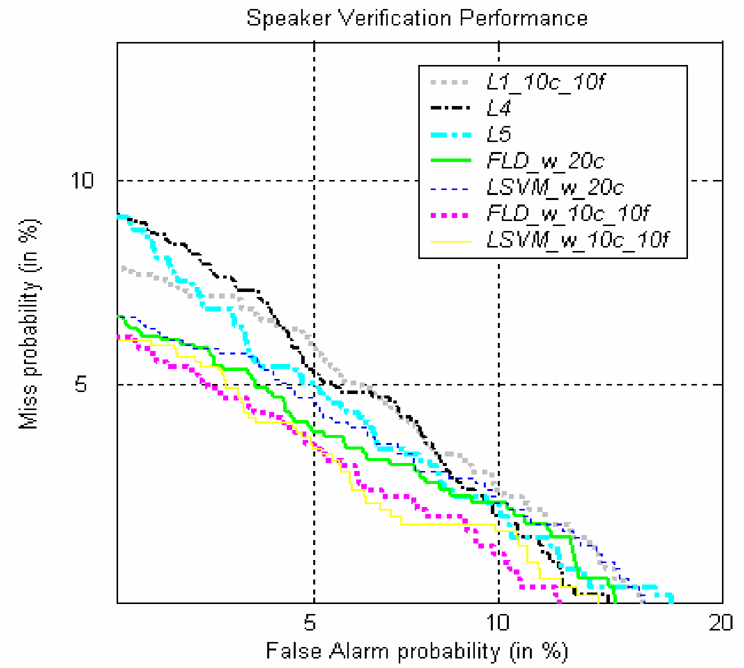


Figure 2. Best baselines vs. our proposed LLR systems based on linear-based classifiers: DET curves for the XM2VTS "Test" subset (B = 20).

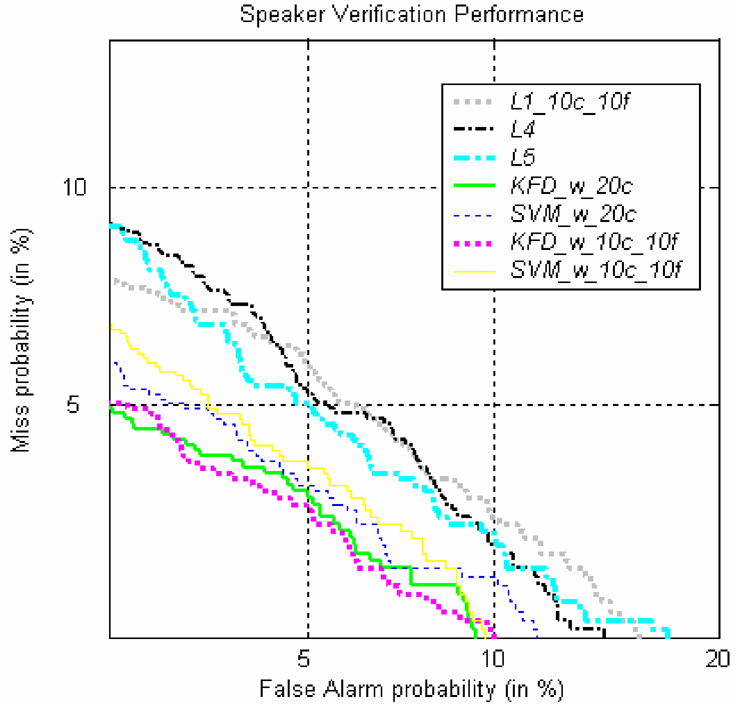


Figure 3. Best baselines vs. our proposed LLR systems based on kernel-based classifiers: DET curves for the XM2VTS "Test" subset (B = 20).

An analysis of the results based on HTER is given in Table 3. For each approach, the decision threshold, θ or b, was used to minimize HTER on the "Evaluation" subset and then applied to the "Test" subset. From Table 3, we observe that all the proposed LLR systems outperform the baseline systems and, for the "Test" subset, a 29.72% relative improvement was achieved by "KFD_w_20c", compared to "L5" – the best baseline system. The advantage of integrating multiple background models in our methods could be the reason why the proposed LLR systems based on the linear SVM ("LSVM_w_20c" and "LSVM_w_10c_10f") outperform "L5", which applied the kernel-based SVM in $L_5(U)$. We also observe that, in the proposed LLR systems, all of the kernel-based methods outperform the linear-based methods.

To analyze the effect of the number of background models, we implemented several proposed LLR systems and baseline systems with $B = 40$. An analysis of the results based on the HTER is given in Table 4. Compared to Table 3, the performance of each system with $B = 40$ is, in general, better than that of its counterpart with $B = 20$, but not always. For instance, "KFD_w_20c_20f" in Table 4 achieved a lower HTER for "Evaluation" but a higher HTER for "Test", compared to "KFD_w_10c_10f" in Table 3. This may be the result of overtraining. However, from Table 4, it is clear that the superiority of the proposed LLR systems over the baseline systems is again demonstrated.

Table 3. HTERs for the XM2VTS "Evaluation" and "Test" subsets (B = 20).

	min HTER for "Evaluation"	HTER for "Test"
*L*1_20c	0.0676	0.0535
*L*1_10c_10f	0.0589	0.0515
*L*2_20c	0.0776	0.0635
*L*3_20c	0.0734	0.0583
*L*4	0.0633	0.0519
*L*5	0.0590	0.0508
FLD_w_20c	0.0459	0.0433
LSVM_w_20c	0.0472	0.0495
FLD_w_10c_10f	0.0468	0.0455
LSVM_w_10c_10f	0.0453	0.0434
KFD_w_20c	0.0247	0.0357
SVM_w_20c	0.0320	0.0414
KFD_w_10c_10f	0.0232	0.0389
SVM_w_10c_10f	0.0310	0.0417

Table 4. HTERs for the XM2VTS "Evaluation" and "Test" subsets (B = 40).

	min HTER for "Evaluation"	HTER for "Test"
*L*1_40c	0.0675	0.0493
*L*1_20c_20f	0.0589	0.0506
*L*2_40c	0.0765	0.0597
*L*3_40c	0.0722	0.0554
KFD_w_40c	0.0074	0.0345
SVM_w_40c	0.0189	0.0386
KFD_w_20c_20f	0.0050	0.0416
SVM_w_20c_20f	0.0192	0.0403

5.2 Evaluation on the ISCSLP2006-SRE Database

We participated in the text-independent speaker verification task of the ISCSLP2006 Speaker Recognition Evaluation (ISCSLP2006-SRE) plan [Chinese Corpus Consortium 2006]. The database contained 800 clients. Each client has one long training utterance, ranging in duration from 21 to 85 seconds, with an average length of 37.06 seconds. In addition, there are 5,933 utterances in the "Test" subset, each of which ranges in duration from 5 seconds to 54 seconds, with an average length of 15.66 seconds. Each test utterance is associated with the client claimed by the speaker, and the task is to judge whether it is true or false. The ratio of true clients to imposters is approximately 1:20. The answer sheet was released after the evaluation finished.

To form the "Evaluation" subset for estimating θ, $\mathbf{w}$, $\mathbf{w}_F$, and b, we extracted some speech from each client's training utterance in the following way. First, we sorted the 800 clients in descending order according to the length of their training utterances. Then, for the first 100 clients, we cut two 4-second segments from the end of each client's training utterance; however, for the remaining 700 clients, we only cut one 4-second segment from the end of each client's training utterance. This yielded 900 (2×100+700) "Evaluation" utterances. In estimating θ, $\mathbf{w}$, $\mathbf{w}_F$, and b, each "Evaluation" utterance served as a client sample for its associated client, but acted as an imposter sample for each of the remaining 799 clients. This yielded 900 client samples and 719,100 (900×799) impostor samples. We used all the client samples and 2,400 randomly-selected impostor samples to estimate $\mathbf{w}_F$ of the kernel-based classifiers. To determine θ or b, we used the 900 client samples and 18,000 randomly-selected impostor samples. This follows the suggestion in the ISCSLP2006-SRE Plan that the ratio of true clients to imposters in the "Test" subset should be approximately 1:20.

The remaining portion of each client's training utterance was used as "Training" to train that client's model through UBM-MAP adaptation [Reynolds 2000]. This was done by first pooling all the speech in "Training" to train a UBM [Reynolds 2000] with 1,024 mixture Gaussian components, and then adapting the mean vectors of the UBM to each client's GMM according to his/her "Training" utterance.

The signal processing front-end was same as that applied in the XM2VTS task.

5.2.1 Experiment Results

The GMM-UBM [Reynolds 2000] system is the current state-of-the-art approach for the text-independent speaker verification task. Thus, in this part, we focus on the performance improvements of our methods over the baseline GMM-UBM system.

As with the GMM-UBM system, we used the fast scoring method [Reynolds 2000] for likelihood ratio computation in the proposed methods. Both the client model λ and the B cohort models were adapted from the UBM Ω. Since the mixture indices were retained after UBM-MAP adaptation, each element of the characteristic vector $\mathbf{x}$ was computed approximately by only considering the C mixture components corresponding to the top C scoring mixtures in the UBM [Reynolds 2000]. In our experiments, the value of C was set to 5.

B was set to 100 in the experiments. We implemented the proposed LLR system in four ways: 1) KFD with Eq. (30) ("KFD_w_100c"), 2) KFD with Eq. (31) ("KFD_w_50c_50f"), 3) SVM with Eq. (30) ("SVM_w_100c"), and 4) SVM with Eq. (31) ("SVM_w_50c_50f"). We compared the proposed systems with the baseline GMM-UBM system and Bengio *et al.*'s system (*L*5). Figure 4 shows the results of experiments conducted on 5,933 "Test" utterances in DET curves. The proposed LLR systems clearly outperform the baseline GMM-UBM system and Bengio *et al.*'s system (*L*5). According to the ISCSLP2006 SRE plan, the performance is measured by the NIST DCF with $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, and $P_{Target} = 0.05$. In each system, the decision threshold, θ or *b*, was selected to minimize the DCF on the "Evaluation" subset, and then applied to the "Test" subset. The minimum DCFs for the "Evaluation" subset and the associated DCFs for the "Test" subset are given in Table 5. We observe that "KFD_w_50c_50f" achieved a 34.08% relative improvement over "GMM-UBM", and a 19.73% relative improvement over "*L*5".

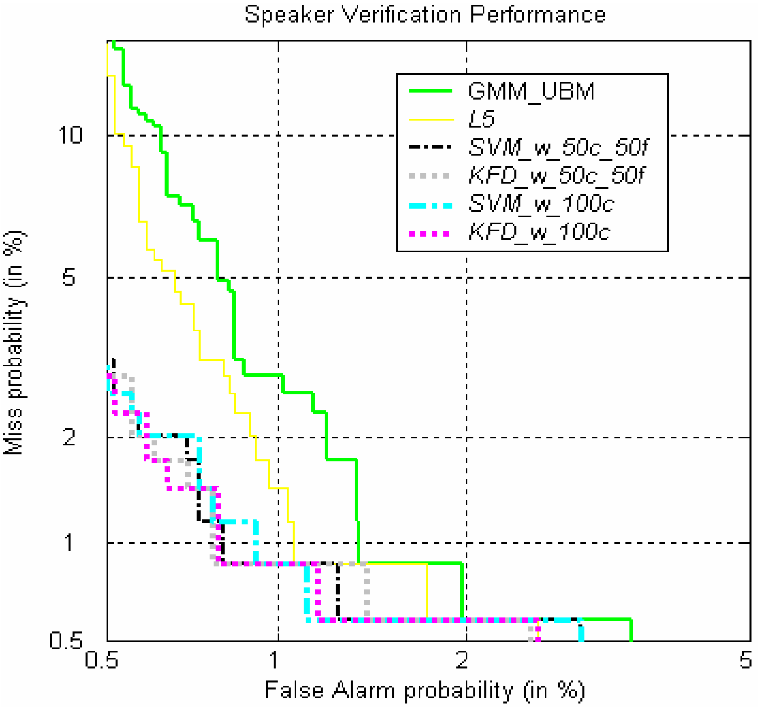


Figure 4. DET curves for the ISCSLP2006-SRE "Test" subset.

Table 5. DCFs for the ISCSLP2006-SRE "Evaluation" and "Test" subsets.

	min DCF for "Evaluation"	DCF for "Test"
GMM-UBM	0.0129	0.0179
*L*5	0.0120	0.0147
KFD_w_50c_50f	0.0067	0.0118
SVM_w_50c_50f	0.0067	0.0123
KFD_w_100c	0.0063	0.0145
SVM_w_100c	0.0076	0.0142

6. Conclusions

We have presented a new LLR measure for speaker verification that improves the characterization of the alternative hypothesis by integrating multiple background models in a more effective and robust way than conventional methods. This new LLR measure is formulated as a non-linear classification problem and solved by using kernel-based classifiers, namely, the Kernel Fisher Discriminant and Support Vector Machine, to optimally separate the LLR samples of the null hypothesis from those of the alternative hypothesis. Experiments, in which the proposed methods were applied to two speaker verification tasks, showed notable improvements in performance over classical LLR-based approaches. Finally, it is worth noting that the proposed methods can be applied to other types of data and hypothesis testing problems.

References

Bengio, S. and J. Mariéthoz, "Learning the Decision Function for Speaker Verification," In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2001, Salt Lake City, USA, pp. 425-428.

Burges, C., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2, 1998, pp. 121-167.

Chinese Corpus Consortium (CCC), "Evaluation Plan for ISCSLP'2006 Special Session on Speaker Recognition," 2006.

Duda, R. O., P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, 2001.

Huang, X., A. Acero and H. W. Hon, *Spoken Language Processing*, Prentics Hall, New Jersey, 2001.

Liu, C. S., H. C. Wang and C. H. Lee, "Speaker Verification Using Normalized Log-Likelihood Score," *IEEE Trans. Speech and Audio Processing*, 4, 1996, pp.56-60.

Luettin, J. and G. Maitre, "Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB)," IDIAP-COM 98-05, IDIAP, 1998.

Martin, A., G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," In *Proceedings of Eurospeech*, 1997, pp. 1895-1898.

Messer, K., J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," In *Proceedings of the 2nd International Conference on Audio and Video-based Biometric Person Authentication*, 1999, Washington D. C., USA, pp. 72-77.

Mika, S., G. Rätsch, J. Weston, B. Schölkopf and K. R. Müller, "Fisher Discriminant Analysis with Kernels," *Neural Networks for Signal Processing IX*, 1999, pp. 41-48.

Reynolds, D. A., "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication*, 17, 1995, pp. 91-108.

Reynolds, D. A., T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, 10, 2000, pp. 19-41.

Rosenberg, A. E., J. Delong, C. H. Lee, B. H. Juang and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," In *Proceedings of International Conference on Spoken Language Processing*, 1992, Banff, Canada, pp. 599-602.

Wan ,V. and S. Renals, "Speaker Verification Using Sequence Discriminant Support Vector Machines," *IEEE Trans. Speech and Audio Processing*, 13(2), 2005, pp. 203-210.

Vapnik, V., *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 273-290

Integrating Complementary Features from Vocal Source and Vocal Tract for Speaker Identification

Nengheng Zheng[*], Tan Lee[*], Ning Wang[*] and P. C. Ching[*]

Abstract

This paper describes a speaker identification system that uses complementary acoustic features derived from the vocal source excitation and the vocal tract system. Conventional speaker recognition systems typically adopt the cepstral coefficients, *e.g.*, Mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC), as the representative features. The cepstral features aim at characterizing the formant structure of the vocal tract system. This study proposes a new feature set, named the wavelet octave coefficients of residues (WOCOR), to characterize the vocal source excitation signal. WOCOR is derived by wavelet transformation of the linear predictive (LP) residual signal and is capable of capturing the spectro-temporal properties of vocal source excitation. WOCOR and MFCC contain complementary information for speaker recognition since they characterize two physiologically distinct components of speech production. The complementary contributions of MFCC and WOCOR in speaker identification are investigated. A confidence measure based score-level fusion technique is proposed to take full advantage of these two complementary features for speaker identification. Experiments show that an identification system using both MFCC and WOCOR significantly outperforms one using MFCC only. In comparison with the identification error rate of 6.8% obtained with MFCC-based system, an error rate of 4.1% is obtained with the proposed confidence measure based integrating system.

Keywords: Speaker Identification, Vocal Source Feature, Vocal Tract Feature, Information Fusion, Confidence Measure

1. Introduction

Speaker recognition is the process of determining a person's identity based on the intrinsic characteristics of his/her voice. In the source-filter model of human speech production, the

[*] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong.
E-mail: nhzheng@ee.cuhk.edu.hk

[Received May 10, 2007; Revised August 9, 2007; Accepted August 12, 2007]

speech signal is modeled as the convolutional output of a vocal source excitation signal and the impulse response of a vocal tract filter system [Rabiner and Schafer 1978]. The most representative vocal tract related acoustic features are the cepstral coefficients, *e.g.*, Mel-frequency cepstral coefficients (MFCC) [Davis and Mermelstein 1980] and linear predictive cepstral coefficients (LPCC) [Furui 1981], which aim at modeling the spectral envelope, or the formant structure of the vocal tract. With the primary goal being identifying different speech sounds, these features are believed to provide pertinent cues for phonetic classification and have been successfully applied to automatic speech recognition [Rabiner and Juang 1993]. At the same time, these features are also implemented in most existing speaker recognition systems [Campbell 1997; Reynolds 2002]. This indicates that MFCC and LPCC features do contain important speaker-specific information, in addition to the intended phonetic information. Ideally, if a large amount of phonetically balanced speech data is available for speaker modeling, the phonetic variability tends to be smoothed out so that speaker-specific aspects can be captured.

The vocal source related features, *e.g.*, pitch and harmonics, on the other hand, characterize the vocal folds' vibration style in speech production and are closely related to the speaker-specific laryngeal system. The spoken contents have less effect on the variation of the vocal source excitation than on that of the vocal tract system [Miller 1963; Childers 1991]. This makes the vocal source derived acoustic features useful for speaker recognition, especially for text-independent cases. However, the usefulness of vocal source information for speaker recognition, although having been investigated in some literature, has not been thoroughly studied, let alone the efficient information retrieving techniques. In this paper, a novel vocal source feature is presented and implemented to supplement the vocal tract features in speaker recognition.

For voiced speech, the source excitation signal is a quasi-periodic glottal waveform, which is generated with quasi-periodic vocal fold vibration. The vibration frequency determines the pitch of voice. It has been shown that temporal pitch variation is useful for speaker recognition [Atal 1972; Sonmez 1998]. The amplitude of pitch harmonics has also been demonstrated to be an effective feature for speaker identification [Imperl *et al.* 1997]. To exploit detailed vocal source information, we need a method of automatically estimating the glottal waveform from the speech signal. This can be done by inverse filtering the speech signal with the vocal tract filter parameters estimated during the glottal closing phase (GCI). In Brookes and Chan [1994], a separately recorded laryngograph signal was used to detect the GCI. In Plumpe *et al.* [1999], a method of automatic GCI detection was proposed and the estimated glottal waveform was represented using the Liljencrants-Fant (LF) model. The model parameters were shown to be useful in speaker identification. However, this method worked well only for the typical voices in which the GCI clearly exists and the estimated glottal waveform can be well explained by the LF model [Plumpe *et al.* 1999].

In linear predictive (LP) modeling of speech signals, the vocal tract system is represented by an all-pole filter. The prediction error, which is named the LP residual signal, contains useful information about the source excitation [Rabiner and Schafer 1978]. In Thevenaz and Hugli [1995], it is shown that the cepstrum of LP residual signal could be used to improve the performance of a text-independent speaker verification system. In He *et al.* [1995] and Chen and Wang [2004], the standard procedures for extracting MFCC and LPCC features were applied to LP residual signals, resulting in a set of residual features for speaker recognition. In Yegnanarayana *et al.* [2005], the speaker information present in LP residual signals was captured using an auto-associative neural network model. Murty and Yegnanarayana [2006] proposed to extract residual phase information by applying Hilbert transform on LP residual signals. The phase features were used to supplement MFCC in speaker recognition.

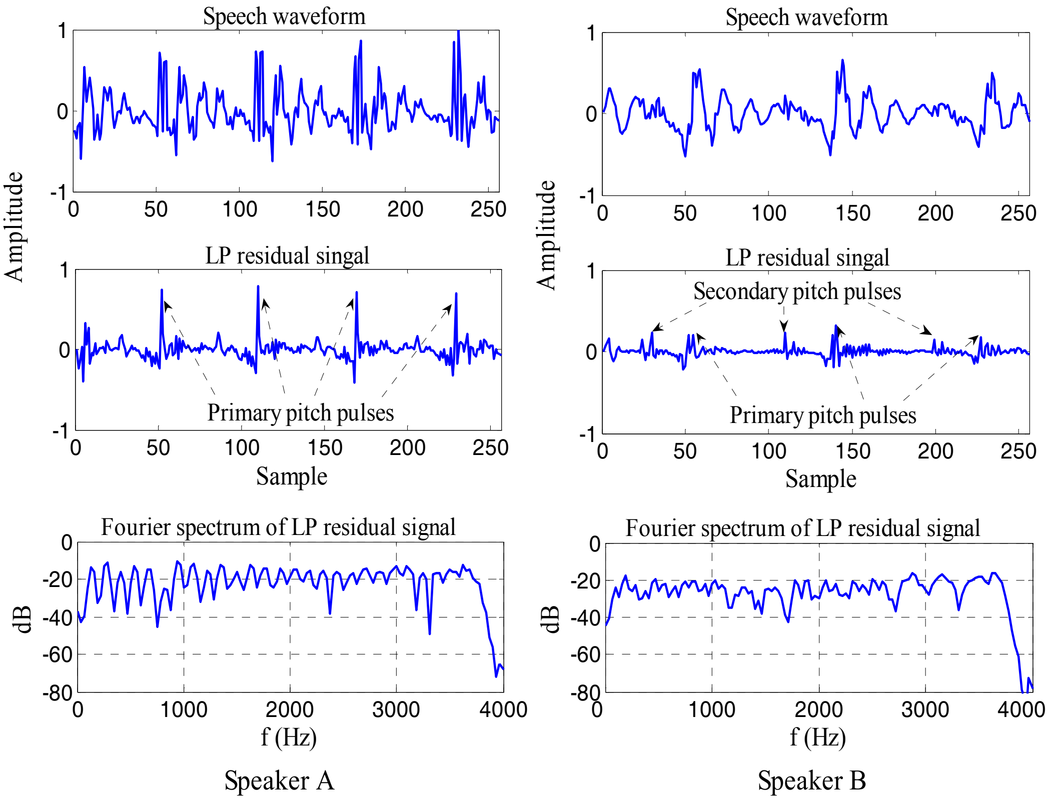


Figure 1. Examples of speech waveforms and LP residual signals of two male speakers. Left: Speaker A; Right: Speaker B; Top to bottom: speech waveforms, LP residual signals and Fourier spectra of LP residual signals

Figure 1 shows the speech waveforms of the vowel /a/ uttered by two different male speakers and the corresponding LP residual signals. There are noticeable differences between the two segments of residual signals. In addition to the difference between their pitch periods, the residual signal of speaker A shows much stronger periodicity than that of speaker B. For speaker B, the magnitudes of the secondary pulses are relatively high. In frequency domain, the Fourier spectra of the two residual signal segments look similar in that they have nearly flat envelopes. Although the harmonic peaks carry speaker-related periodicity information, the useful temporal information, *i.e.*, the amplitudes and the time locations of pitch pulses, are not represented in the Fourier spectra. To characterize the time-frequency characteristics of the pitch pulses, wavelet transform is more appropriate than the short-time Fourier transform.

This paper describes a novel feature extraction technique based on time-frequency analysis of the LP residual signal. The new feature parameters, called wavelet octave coefficients of residues (WOCOR), are generated by applying pitch-synchronous wavelet transform to the residual signal [Zheng *et al.* 2004]. The WOCOR features contain useful information for speaker characterization and recognition. More importantly, WOCOR and MFCC carry different speaker-specific information since they characterize two physiologically distinct components in speech production. As a result, combining these two complementary features will result in higher recognition performance than using only one set of features.

The performance of the information fusion system, however, is highly dependant on the effectiveness of the fusion technique implemented. In multi-modal biometric authentication systems, the reliability of authentication decisions from different classifiers may vary significantly in different tests. Therefore, it is very important to apply an efficient fusion technique to maximize the benefit through the information fusion. A number of information fusion techniques have been proposed for biometrics systems [Garcia-Romero *et al.* 2004; Ross *et al.* 2001; Toh and Tau 2005]. Generally, the information fusion can be done at: (i) feature level, (ii) score level, or (iii) decision level. This paper proposes a score level fusion technique for combining MFCC and WOCOR for speaker identification. Score level fusion is preferred because the matching scores are easily obtained and contain sufficient information for distinguishing different speakers. A confidence measure, which measures the confidence of MFCC in identification decision in comparison to that of WOCOR, is adopted as the fusion weight in each individual identification trial. The confidence measure provides an optimized fusion score by giving more weight to the feature of higher confidence in correct identification. The effectiveness of the proposed information fusion system is demonstrated by a set or speaker identification experiments.

The rest of this paper is organized as follows. Section 2 describes the feature extraction procedures for WOCOR and briefly reviews the MFCC feature extraction procedures. Section

3 demonstrates the usefulness of WOCOR in speaker identification and the complementary contributions of WOCOR and MFCC in speaker identification. Section 4 presents the confidence measure based score-level fusion technique for integrating MFCC and WOCOR for speaker identification. Some analysis of the identification results is presented in Section 5, which further elaborates the complementarity of MFCC and WOCOR in speaker recognition and the superiority of the proposed confidence measure based fusion technique over the fixed-weight fusion. Conclusions are given in Section 6.

2. Vocal Source and Vocal Tract Features

2.1 Vocal Source Features: WOCOR

As illustrated in Figure 1, Fourier spectrum is not good at characterizing the time-frequency properties of the pitch pulses in the residual signal. Wavelet transform has been well known to be an effective method for transient signal representation. Therefore, the proposed WOCOR feature extraction is based on wavelet transform, rather than Fourier transform, of the residual signal. The process of extracting the WOCOR features is formulated in the following steps:

1) *Voicing decision and pitch extraction.* Voicing status decision and pitch extraction are done with Talkin's Robust Algorithm for Pitch Tracking [Talkin 1995]. Only voiced speech is retained for subsequent processing. In the source-filter model, the excitation signal for unvoiced speech can be approximated as random noise [Rabiner and Schafer 1978]. We believe that such noise-like signals carry relatively little speaker-specific information.

2) *LP inverse filtering.* The voiced speech is divided into non-overlapping frames of 30 ms long. The LP residual signal $e(n)$ is obtained from each frame by inverse filtering the speech signal $s(n)$, *i.e.*,

$$e(n) = s(n) - \sum_{k=1}^{12} a_k s(n-k) \tag{1}$$

where the LP filter coefficients a_k are computed using the autocorrelation method [Rabiner and Schafer 1978]. To reduce intra-speaker variation, the amplitude of the residual signal within each voiced segment is normalized to the range [-1, 1].

3) *Pitch-synchronous windowing.* Based on the pitch periods estimated in Step 1, pitch pulses in the residual signal are located by detecting the maximum amplitude within each pitch period. For each pitch pulse, pitch-synchronous wavelet analysis is applied with a Hamming window of two pitch periods long. Let t_{i-1}, t_i and t_{i+1} denote the locations of three successive pitch pulses. The analysis window for the pitch pulse at t_i spans from t_{i-1} to t_{i+1}, as illustrated in Figure 2. The windowed residual signal is denoted as $e_h(n)$.

4) *Wavelet transform of residual signal.* The wavelet transform of $e_h(n)$ is computed as:

$$w(a,b) = \frac{1}{\sqrt{|a|}} \sum_n e_h(n) \Psi^* \left(\frac{n-b}{a} \right) \tag{2}$$

where $a = \{2^k \,|k = 1,2,\cdots,K\}$ and $b = 1,2,\cdots,N$, and N is the window length. $\Psi^*(n)$ is the conjugate of the 4th-order Daubechies wavelet basis function $\Psi(n)$. a and b are the scaling parameter and the translation parameter, respectively [Daubechies 1992]. In this case, the LP residual signal is analyzed in K octave sub-bands. For a specific sub-band, the time-varying characteristics within the analysis window are measured as b changes.

5) *Generation of WOCOR feature parameters.* We have K octave groups of wavelet coefficients, *i.e.*,

$$W_k = \left\{ w\left(2^k, b\right) \;\middle|\; b = 1,2,\cdots,N \right\}, \quad k = 1,2,\cdots,K \tag{3}$$

To retain the temporal information, each octave group of coefficients is divided evenly into M sub-groups, *i.e.*,

$$W_k^M(m) = \left\{ w\left(2^k, b\right) \;\middle|\; b \in \left(\frac{(m-1)N}{M}, \frac{mN}{M} \right] \right\}, \quad m = 1,2,\cdots,M \tag{4}$$

where M is the number of sub-groups. The 2-norm of each sub-group of coefficients is computed to be one of the feature parameters. As a result, the complete feature vector is composed of $K \cdot M$ parameters as follows,

$$\text{WOCOR} = \left\{ \left\| W_k^M(m) \right\| \;\middle|\; \begin{array}{l} m = 1,2,\cdots,M \\ k = 1,2,\cdots,K \end{array} \right\} \tag{5}$$

where $\|\bullet\|$ denotes the 2-norm operation.

Figure 2 illustrates the extraction of WOCOR features from a pitch-synchronous segment of residual signal. It can be seen that, with different values of *k*, the signal is analyzed with different time-frequency resolutions. The time-frequency properties of the signal in each sub-band are characterized by the wavelet coefficients. In this research, we are interested in telephone speech with the frequency band of 300 - 3400 Hz. To cover this range, we set $K = 4$ and the four frequency sub-bands at different octave levels are defined accordingly: 2000 - 4000 Hz (W_1), 1000 – 2000 Hz (W_2), 500 - 1000 Hz (W_3), and 250 - 500 Hz (W_4). The parameter M determines the temporal resolution attained by the WOCOR parameters. If $M = 1$, all the coefficients of a sub-band are combined into a single feature parameter, and no temporal information is retained. On the other hand, if a large M is used, such that each coefficient acts as an individual feature parameter, a lot of unnecessary temporal details are included and the feature vector tends to be noisy and less discriminative. A low feature dimension is also desirable for effective statistical modeling. In Section 3.3, the effect of M on recognition performance will be investigated experimentally.

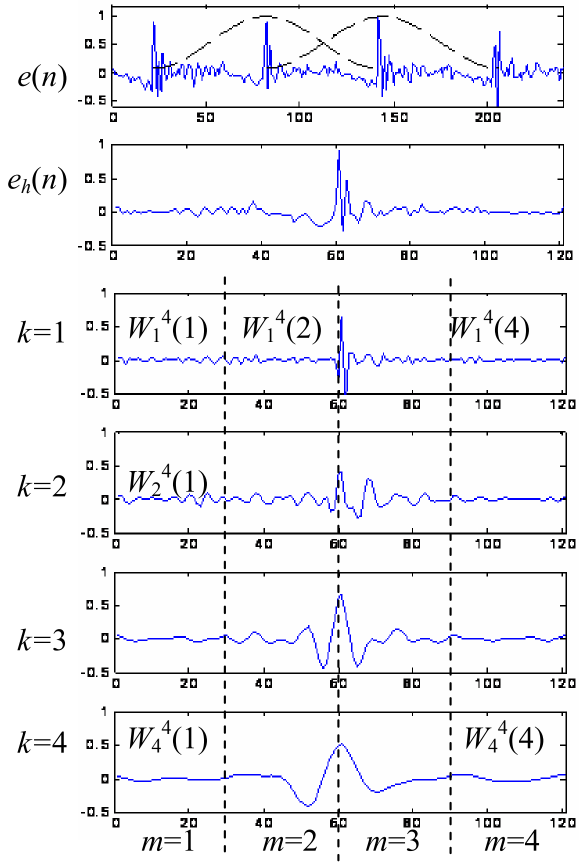


Figure 2. Extraction of WOCOR features from a pitch-synchronous segment of LP residual signal. Here K = 4 and M = 4

To summarize, given a speech utterance, a sequence of WOCOR feature vectors is obtained by pitch-synchronous wavelet transform of the LP residual signal. The WOCOR features are expected to capture spectro-temporal characteristics of the residual signal, which is useful for speaker characterization and recognition.

2.2 Vocal Tract Features: MFCC

The MFCC features have been widely used for speech and speaker recognition. In this study, we use the standard procedures of extracting MFCC on a short-time frame basis as described below [Davis and Mermelstein 1980]:

1) Short-time Fourier transform is applied every 10 ms with 30-ms Hamming window.

2) The magnitude spectrum is warped with a Mel-scale filter bank that consists of 26 filters,

which emulates the frequency resolution of human auditory system.

3) The log-energy of each filter output is computed.

4) Discrete cosine transform (DCT) is applied to the filter-bank output to produce the cepstral coefficients.

The MFCC feature vector has 39 components, including the first 12 cepstral coefficients, the log energy, as well as their first and second order time derivatives.

Aiming at characterizing two physiologically distinct components in speech production, WOCOR and MFCC contain complementary information for speaker discrimination. The effectiveness of WOCOR and its complementarity to MFCC for speaker recognition will be investigated in the following sections.

3. Experiments

3.1 Speaker Identification System

Figure 3 gives the block diagram of the speaker identification system using MFCC and WOCOR. In the pre-processing stage, the speech signal is first pre-emphasized with a first order filter $H(z)=1-0.97z^{-1}$. Then energy-based voice activity detection (VAD) technique is applied to remove the silent portion. The speech signal is passed through for MFCC and WOCOR generation, respectively. For each feature set, speaker models are trained with the UBM-GMM technique [Reynolds *et al.* 2000] in the training stage. A universal background model (UBM) is first trained using the training data from all speakers. Then a Gaussian mixture model (GMM) is adapted from the UBM for each speaker using the respective training data. In the test stage, for each identification trial, likelihoods scores of the two feature sets are first computed and then a score-level fusion is implemented, *i.e.*,

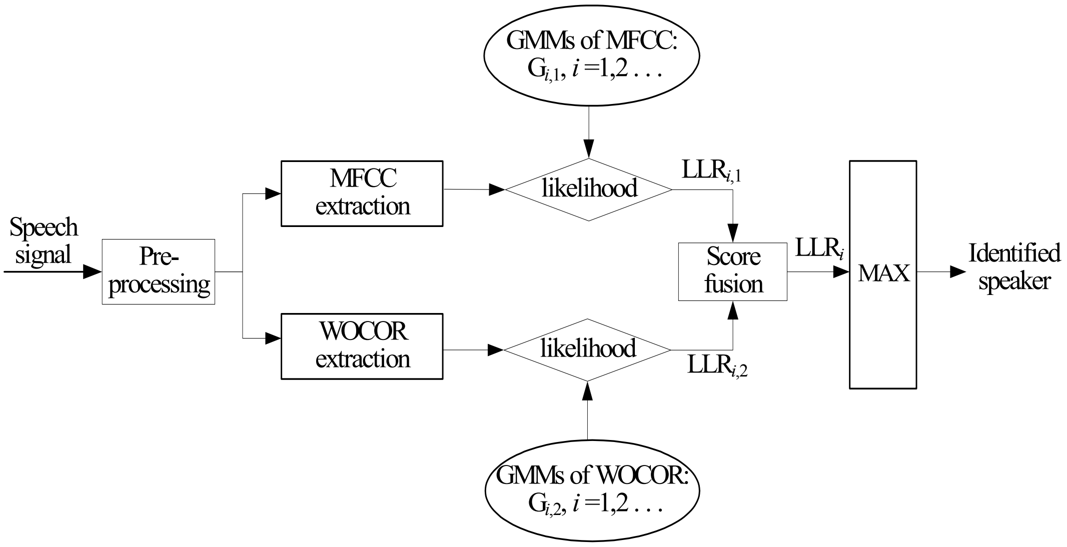


Figure 3. Block diagram of the speaker identification system using MFCC and WOCOR

$$\mathrm{LLR}_i = \mathbf{f}(\mathrm{LLR}_{i,1}, \mathrm{LLR}_{i,2}),\ i = 1, 2, \cdots, N \tag{6}$$

where $\mathrm{LLR}_{i,1}$ and $\mathrm{LLR}_{i,2}$ are likelihood scores obtained from MFCC and WOCOR, respectively, $\mathbf{f}$ is the combination function and N is the number of speakers. Although in real application, the test utterances could come from the unregistered impostors. In this study, we only deal with the closed-set speaker identification. That is, all the test utterances must come from one of the 50 male speakers. The one whose models give the highest matching score is marked as the identified speaker.

3.2 Speech Databases: CU2C

CU2C is a continuous speech database of Cantonese developed at the Chinese University of Hong Kong [Zheng *et al.* 2005]. Cantonese is one of the most popular Chinese dialects and is spoken by tens of millions of people in southern China. CU2C was designed to facilitate general speaker recognition research. It contains parallel utterances collected over fixed-line telephone channel and desktop computer microphones. The spoken contents include Hong Kong personal identity numbers, randomly generated digit strings, and phonetically balanced sentences. In this study, the speaker identification experiments are conducted on the sentence subset of the male speakers. There are 50 male speakers, each having 18 sessions of speech data with 10 utterances in each session. The first 4 sessions are used for training the speaker models. Sessions 5 to 8 are used as development data for training the weighting parameters for the score level fusion of MFCC and WOCOR. The last 10 sessions are used as the evaluation data, and there are totally 5000 identification trials (50 speakers, 100 trials per speaker). All the utterances are text-independent telephone speech with matched training and testing conditions (the same handset and fixed line telephone network). The speech data were sampled at 8 KHz and encoded by 8-bit μ-law encoding. The speech data of each speaker are collected over 4 to 9 months with the minimum inter-session interval of 1 week. Therefore, the challenge of the long-term intra-speaker variation for speaker recognition can be addressed by the database.

3.3 Determining the Parameter *M* for WOCOR

As discussed earlier, the value of M controls the size of the WOCOR feature vector and how much temporal detail can be captured. First, we compare the performance of WOCOR with different values of M. Figure 4 shows the identification error rate (IDER) of WOCOR in which M varies from 1 to 6. The identification error rate is defined as:

$$\mathrm{IDER} = \frac{\text{Number of incorrect identification trials}}{\text{Number of identification trials}} \times 100\% \tag{7}$$

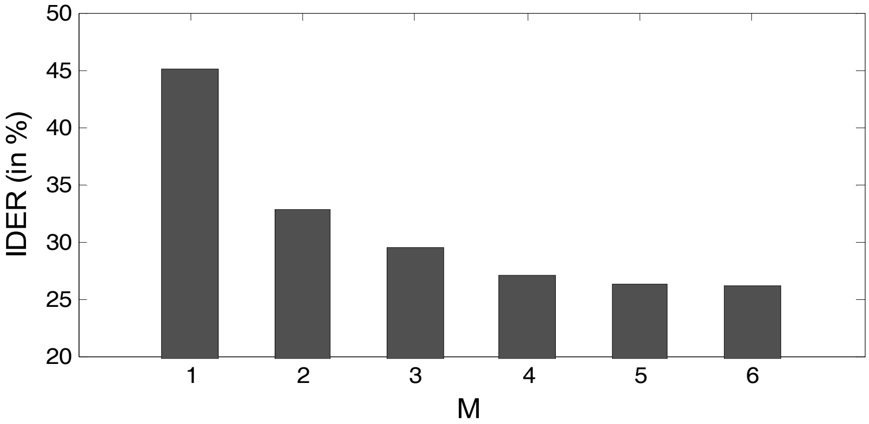


Figure 4. The speaker identification results of WOCOR for different values of M

It is clear that WOCOR in general provide a certain degree of speaker discrimination power. For $M = 1$, *i.e.*, no temporal detail is captured and the feature vector has only 4 components, an IDER of 45.1% is achieved. With M increasing from 1 to 4, the IDER is significantly reduced to only 27.0%. For $M > 4$, the improvement becomes less noticeable. Therefore, in the following experiments, we will use WOCOR with $M = 4$, which consists of 16 feature components.

3.4 Wavelet vs. Fourier Transform of LP Residual Signal

To demonstrate the superiority of wavelet transform over Fourier transform for feature extraction from the LP residual signal, we compare the speaker identification performances of WOCOR and the Fourier spectrum-based vocal source features. To do so, we apply the MFCC feature extraction process on the LP residual signal to generate another set of vocal source features, noted as $\mathrm{MFCC_{res}}$. Speaker identification experiments with WOCOR and $\mathrm{MFCC_{res}}$ result in IDERs of 27.0% and 52.0%, respectively. That is, WOCOR significantly outperforms $\mathrm{MFCC_{res}}$. This is reasonable because MFCC focuses on extracting the spectral envelope-related features, and, as given in Fig. 1, spectral envelopes of LP residual signals are almost the same for different speakers. On the other hand, WOCOR tries to capture the spectro-temporal information in the residual signals, which is quite different between speakers.

3.5 Speaker Identification Results

We evaluate the speaker identification performances of MFCC and WOCOR individually. In addition, we evaluate the system with both MFCC and WOCOR, using the same evaluation data described in Section 3.2 for all three performance evaluations. In this case, information fusion is performed as a score-level linear fusion, *i.e.*,

$$\mathrm{LLR}_i = w_t \mathrm{LLR}_{i,1} + (1 - w_t)\mathrm{LLR}_{i,2} \tag{8}$$

The fusion weight w_t is experimentally determined using the development data set. That is, w_t is varied from 0 to 1, and the value giving the smallest IDER is selected for the evaluation trials. Figure 5 shows IDER vs. w_t curve with the development data. As illustrated, the best performance is achieved at around $w_t = 0.80$. Actually, the identification performance is not very sensitive to w_t at around $w_t = 0.80$. The performances of MFCC- and WOCOR-based systems and the information fusion system with $w_t = 0.80$ are evaluated over the evaluation data and the results are as given in Table 1. As shown, the MFCC-based speaker identification system significantly outperforms the WOCOR system. It is noted that, despite the performance difference, the two approaches make complementary decisions in many cases, which will be further elaborated in Section 5, and the combining system has superior performance over that using MFCC only. The IDER is reduced from 6.8% to 4.7%, a relative improvement of about 30%.

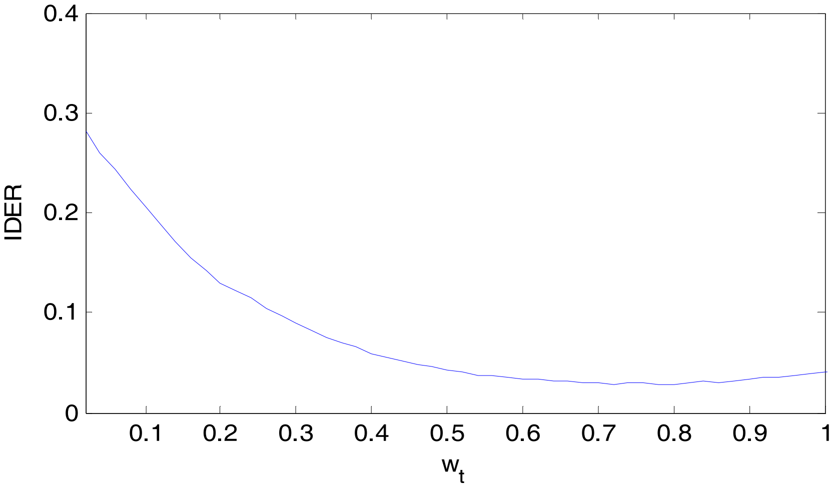


Figure 5. Speaker identification performance with various w_t

Table 1. Speaker identification performances

Systems	IDER (in %)
WOCOR	27.0
MFCC	6.8
MFCC+WOCOR	4.7

4. Information Fusion with Confidence Measure

While information fusion with a pre-defined fusion weight as given in (8) can improve identification performance, it does not necessarily provide the best result. Fixed weight is unable to cover explicitly the different performance levels of MFCC and WOCOR for individual identification trials. As a result, for some cases, although one of the features gives the correct decision, the fused score may not necessarily result in correct decision. For example, consider four types of identification trials as given in Table 2, in which MFCC and WOCOR give different contributions to speaker identification, and the info-fusion as (8) results in different decisions as well. In Type I and II trials, MFCC gives incorrect decisions while WOCOR gives correct decisions. The combined system makes correct decisions in Type I trials while making incorrect decisions in Type II trials. In Type III and IV trials, MFCC gives correct decisions while WOCOR gives incorrect decisions, and the combined system makes correct decisions in Type III trials while producing incorrect decisions in Type IV trials. To avoid the undesired outputs in Type II and IV trials, an ideal solution should be capable of distinguishing these four types of trials and give null weight to MFCC in Type I and II trials and null weight to WOCOR in Type III and IV trials. Although such an ideal solution is not available in real-world applications, we propose to apply a confidence measure based fusion method, which adopts varying weight in individual trials and avoids most of the identification errors introduced by information fusion.

Table 2. Different contributions of MFCC and WOCOR in four types of identification trials

	Type I	Type II	Type III	Type IV
MFCC	incorrect	incorrect	correct	correct
WOCOR	correct	correct	incorrect	incorrect
MFCC+WOCOR	correct	incorrect	correct	incorrect

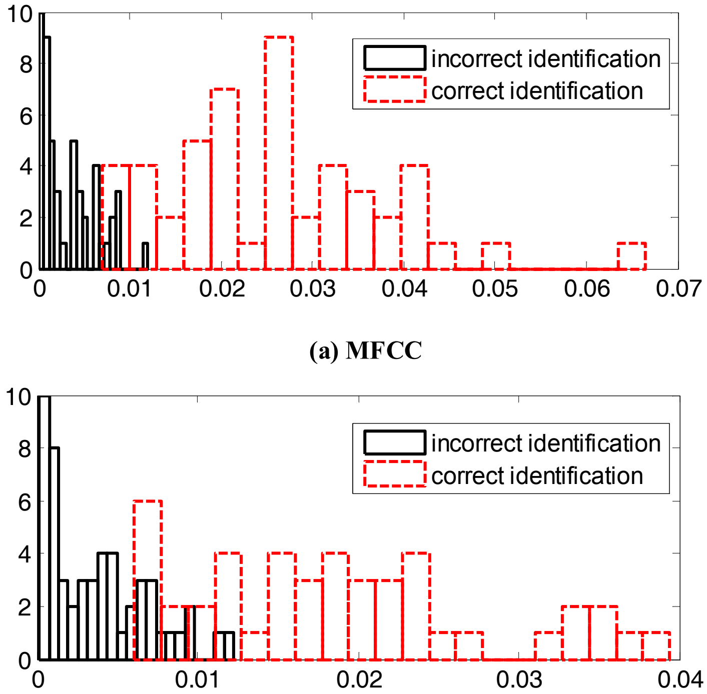
4.1 Speaker Discrimination Power

Analysis of the matching scores shows that, generally, in a correct identification, the difference of the scores between the identified speaker and the closest competitor is relatively larger than that in an incorrect identification. The score difference can therefore be adopted for measuring the speaker discrimination power, *i.e.*,

$$D = \frac{\max_i \{\mathrm{LLR}_i\} - \sec ond \max_i \{\mathrm{LLR}_i\}}{\left| \max_i \{\mathrm{LLR}_i\} \right|} \qquad (9)$$

where LLR_i is the likelihood score of the *i*-th speaker. The normalization of the difference over $\left|\max_i\{\mathrm{LLR}_i\}\right|$ aims to equalize the dynamic ranges of *D* for different features.

Figure 6 shows the histograms of *D* for MFCC and WOCOR. It is clear that, for both features, a correct identification is generally associated with a larger *D* than an incorrect identification. Therefore, a larger *D* implies that the corresponding feature has higher confidence for speaker identification. Obviously, it is desirable to take into account *D* for score fusion in each identification trial instead of using the fixed weight.



(a) MFCC

(b) WOCOR

Figure 6. Histogram of speaker discrimination power D of MFCC and WOCOR

4.2 Confidence Measure Based Score Fusion

Although the optimal method of combining the scores from MFCC and WOCOR with the knowledge of the discrimination power is not known, the relative discrimination power of MFCC and WOCOR can be considered as a confidence measure, with which a better fusion weight can be derived to improve the identification performance. In each identification trial, the confidence measure is defined the discrimination ratio of the two features, *i.e.*,

$$\mathrm{CM} = \left|D_1 / D_2\right| \qquad (10)$$

where D_1 and D_2 are the speaker discrimination power of MFCC and WOCOR, respectively. A larger CM implies that the MFCC-based system has a higher confidence in giving correct identification result than the WOCOR-based one. Then, the fusion weight for the specific identification trial is derived as:

$$w_{CM} = -\log \frac{1}{1+e^{-\alpha\cdot(\mathrm{CM}-\beta)}} \tag{11}$$

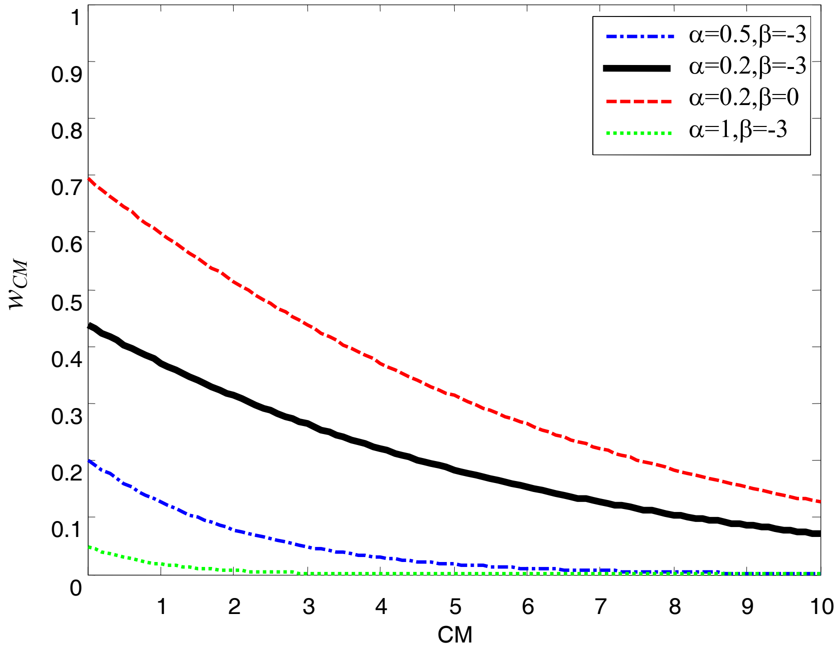


Figure 7. Mapping contours from CM to w_{CM}

where α and β control the slope of the mapping contour from CM to w_{CM}, as illustrated in Figure 7. The solid line curve in Figure 7 is used in this study. The corresponding parameters $\alpha = 0.2, \beta = -3$ are trained using the development data.

Score-level fusion based on CM is then carried out according to:

$$\mathrm{LLR}_i = \mathrm{LLR}_{i,1} + w_{CM}\mathrm{LLR}_{i,2} \tag{12}$$

With w_{CM}, the fused score combines better weighted likelihoods obtained from MFCC and WOCOR in each individual trial based on the contributions of the respective features in that trial.

As illustrated in Figure 7, when CM increases, w_{CM} becomes very small, and the decision will not be heavily affected by WOCOR. On the other hand, a small CM corresponds to a large w_{CM}, which means more impact from WOCOR.

As shown in Table 3, the CM-based score level fusion leads to a further performance improvement over the fixed-weight fusion. In summary, the IDERs attained with WOCOR and MFCC, in conjunction with the two methods of score fusion are 27.0%, 6.8%, 4.7%, and 4.1%, respectively.

Table 3. Speaker identification performances

Systems	IDER (in %)
WOCOR	27.0
MFCC	6.8
Fixed-weight fusion	4.7
Fusion with CM	4.1

5. Analysis of the Identification Results

Table 4 elaborates how the integration of the two complementary features affects the identification performances. The identification trials are divided into 4 subsets according to the performances of MFCC and WOCOR: (i) correct identification with both MFCC and WOCOR (McWc), (ii) incorrect identification with both MFCC and WOCOR (MiWi), (iii) incorrect identification with MFCC while correct identification with WOCOR (MiWc), and (iv) correct identification with MFCC while incorrect identification with WOCOR (McWi). Among the 5000 identification trials, there are 3328, 244, 95 and 1333 trials for these 4 subsets, respectively. The number of identification errors with MFCC, WOCOR and the integrated systems within each subset are given in the table.

Table 4. Number of errors of 4 identification subsets by different systems

Subsets	McWc	MiWi	MiWc	McWi
Number of trials	3328	244	95	1333
MFCC	0	244	95	0
WOCOR	0	244	0	1333
Fixed weight fusion	0	163	7	65
Fusion with CM	0	167	19	19

We are only interested in the last 3 subsets, which have errors with at least one kind of features. For the MiWi subset, although both MFCC and WOCOR give incorrect identification results, the combined system gives correct results for some trials. For example, the number of identification errors is reduced from 244 to 163 with the fixed weight fusion and to 167 with the CM-based fusion. That is, about one third of the errors have been corrected.

Table 5 gives an example demonstrating how the score fusion can give correct result even though both MFCC and WOCOR give error results. In this example, the true speaker is S5. It is shown that although S5 only ranks at the 6th and the 2nd with MFCC and WOCOR, respectively, in both integrating systems, it ranks at the first and therefore is correct identified.

The results of the two one-error identification subsets McWi and MiWc in Table 4 demonstrate the superiority of the CM-based score fusion over the fixed-weight fusion. For the fixed-weight fusion system, although the number of errors in the MiWc subset is significantly reduced from 95 to 7, there are 65 errors introduced to the McWi subset, which have been correctly identified with MFCC only. For the CM-based system, the number of this kind of newly introduced errors is significantly reduced to 19, with only a slight increase in errors in MiWi and MiWc subsets. As a whole, the number of total identification errors is reduced from 339 with MFCC only to 235 with fixed-weight fusion, and further reduced to 205 with CM-based fusion.

Table 5. Ranking the speaker scores in an identification trial.

Rank	MFCC	WOCOR	Fixed weight fusion	Fusion with CM
1	S7: -1.7718	S34:1.5732	**S5:-0.4364**	**S5:-1.0903**
2	S27:-1.7718	**S5:1.5730**	S27:-0.4445	S27:-1.0977
3	S10:-1.7722	S48:1.5640	S34:-0.4446	S7: -1.0984
4	S42:-1.7743	S35:1.5620	S41:-0.4448	S10:-1.1000
5	S1: -1.7756	S39:1.5619	S46:-0.4448	S41:-1.1005
6	**S5:-1.7760**	S46:1.5510	S7: -0.4452	S46:-1.1015
7	S41:-1.7788	S41:1.5561	S10:-0.4465	S42:-1.1027

6. Conclusions

This paper presents a novel feature extraction technique to generate the vocal source related acoustic features from the LP residual signal. We have shown that the proposed WOCOR features contain speaker-specific information for speaker recognition applications. The WOCOR features provide additional information to the conventional MFCC features in speaker recognition. This complementarity is exploited by applying a novel confidence measure based score fusion technique which gives a much improved overall speaker identification accuracy. In comparison with the identification error rate of 6.8% obtained with MFCC only, an error rate of 4.1% is obtained with the proposed information fusion system. That is a relative improvement of 40%.

Acknowledgements

This research was supported partially by an Earmarked Research Grant (Ref. CUHK 4236/04E) and a Central Allocation Grant (Ref. CUHK1/02C) from the Hong Kong Research Grants Council.

References

Atal, B. S., "Automatic speaker recognition based on pitch contours," *Journal of the Acoustical Society of America*, 52(6), 1972, pp. 1687-1697.

Brookes, D. M. and D. S. F. Chan, "Speaker characteristics from a glottal airflow model using robust inverse filtering," *Proceedings of Institute of Acoustics*, 16(5), 1994, pp. 501-508.

Campbell, J. P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, 85(9), 1997, pp. 1437-1462.

Chen, S.-H. and H.-C. Wang, "Improvement of speaker recognition by combining residual and prosodic features with acoustic features," In *Proceedings of 29th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. 93-96.

Childers, D. G. and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, 90(5), 1991, pp. 2394-2410.

Daubechies, I., *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

Davis, S. B. and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 1980, pp. 357-366.

Furui, S., "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 1981, pp. 254 - 272.

Garcia-Romero, D., J. Fierrez-Aguilar, J. Gonzalez-Rodriguez and J. Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," In *ESCA Workshop on Speaker and Language Recognition, Odyssey*, 2004, pp. 105-110.

He, J., L. Liu and G. Palm, "On the use of features from prediction residual signals in speaker identification," In *Proceedings of Eurospeech*, 1995, pp. 313-316.

Imperl, B., Z. Kacic and B. Horvat, "A study of harmonic features for speaker recognition," *Speech Communication*, 22(4), 1997, pp. 385-402.

Miller, J. E. and M. V. Mathews, "Investigation of the glottal waveshape by automatic inverse filtering," *Journal of the Acoustical Society of America*, 35, 1963 pp.1876.

Murty, K. S. and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, 13(1), 2006, pp. 52-55.

Plumpe, M. D., T. F. Quatieri and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, 7(5), 1999, pp. 569-585.

Rabiner, L. R. and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

Rabiner, L. R. and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

Reynolds, D. A., "An overview of automatic speaker recognition technology," In *Proceedings of 27th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 4072-4075.

Reynolds, D. A., T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10(1-3), 2000, pp. 19-41.

Ross, A., A. Jain and J.-Z. Qian, "Information fusion in biometrics," In *Proceedings of 3rd International Conference on Audio- and Video-Based Person Authentication*, 2001, pp. 354-359.

Sonmez, K., E. Shriberg, L. Heck and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," In *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 3189-3192.

Talkin, D., "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, ed. By W. B. Kleijn and K. K. Paliwal, Elsevier, 1995.

Thevenaz, P. and H. Hugli, "Usefulness of the LPC residue in text-independent speaker verification," *Speech Communication*, 17(1-2), 1995, pp. 145-157.

Toh, K.-A. and W.-Y. Yau, "Fingerprint and speaker verification decisions fusion using a functional link network," *IEEE Transactions on System, Man and Cybernetics B*, 35(3), 2005, pp. 357-370.

Yegnanarayana, B., K. S. Reddy and S. P. Kishore, "Source and system features for speaker recognition using AANN models," In *Proceedings of 26th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 409-413.

Zheng, N., P. C. Ching and T. Lee, "Time frequency analysis of vocal source signal for speaker recognition," In *Proceedings of International Conference on Spoken Language Processing*, 2004, pp. 2333-2336.

Zheng, N., C. Qin, T. Lee and P. C. Ching, "CU2C: A dual-condition Cantonese speech database for speaker recognition application," In *Proceedings of Oriental-COCOSDA*, 2005, pp. 67-72.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 291-302

Performance of Discriminative HMM Training in Noise

Jun Du*,+, Peng Liu+, Frank K. Soong+, Jian-Lai Zhou+, and Ren-Hua Wang*

Abstract

In this study, discriminative HMM training and its performance are investigated in both clean and noisy environments. Recognition error is defined at string, word, phone, and acoustic levels and treated in a unified framework in discriminative training. With an acoustic level, high-resolution error measurement, a discriminative criterion of minimum divergence (MD) is proposed. Using speaker-independent, continuous digit databases, Aurora2, the recognition performance of recognizers, which are trained in terms of different error measures and different training modes, is evaluated under various noise and SNR conditions. Experimental results show that discriminatively trained models perform better than the maximum likelihood baseline systems. Specifically, in MWE and MD training, relative error reductions of 13.71% and 17.62% are obtained with multi-training on Aurora2, respectively. Moreover, compared with ML training, MD training becomes more effective as the SNR increases.

Keywords: Noise Robustness, Minimum Divergence, Minimum Word Error, Discriminative Training

1. Introduction

With the progress of Automatic Speech Recognition (ASR), noise robustness of speech recognizers attracts more and more attention for practical recognition systems. Various noise robust technologies can be grouped into three classes: 1. Feature domain approaches, which aim at noise resistant features, *e.g.*, speech enhancement, feature compensation or transformation methods [Gong 1995]; 2. Model domain approaches, *e.g.*, Hidden Markov

* University of Science and Technology of China, Hefei, P. R. China, 230027
Tel: +86-551-3601363-806 Fax: +86-551-3601363-807
E-mail: unuedjwj@ustc.edu; rhw@ustc.edu.cn
+ Microsoft Research Asia, Beijing, P. R. China, 100080
E-mail: {pengliu, frankkps, jlzhou }@microsoft.com

[Received March 31, 2007; Revised August 14, 2007; Accepted August 17, 2007]

Model (HMM) decompensation [Varga *et al.* 1990], Parallel Model Combination (PMC) [Gales *et al*. 1994], which aim at modeling the distortion of features in noisy environments directly; 3. Hybrid approaches.

In the past decade, discriminative training has been shown quite effective in reducing word error rates of HMM based ASR systems in a clean environment. In the first stage, sentence level discriminative training criteria, including Maximum Mutual Information (MMI) [Schluter 2000; Valtchev *et al*. 1997] and Minimum Classification Error (MCE) [Juang *et al*. 1997], were proposed and proven effective. Recently, new criteria such as Minimum Word Error (MWE) and Minimum Phone Error (MPE) [Povey 2004], which are based on fine error analysis at word or phone level, have achieved further improvement in recognition performance.

In [Ohkura *et al*. 1993; Meyer *et al*. 2001; Laurila *et al*. 1998], noise robustness investigation on sentence level discriminative criteria such as MCE, Corrective Training (CT) is reported. Hence, we give a more complete investigation of noise robustness for general minimum error training.

From a unified view of error minimization, the major difference between MCE, MWE and MPE is the error definition. String based MCE is based upon minimizing sentence error rate, while MWE is based on word error rate, which is more consistent with the popular metric used in evaluating ASR systems. Hence, the latter yields a better word error rate, at least on the training set [Povey 2004]. However, MPE performs slightly but universally better than MWE on the testing set [Povey 2004]. The success of MPE might be explained as follows: when refining acoustic models in discriminative training, it makes more sense to define errors in a more granular form of acoustic similarity. However, error definition at phone label level is only a rough approximation of acoustic similarity.

Based on the analysis above, we have proposed using acoustic dissimilarity to measure errors [Du *et al.* 2006]. As acoustic behavior of speech units is characterized by HMMs, by measuring Kullback-Leibler Divergence (KLD) [Kullback *et al.* 1951] between two given HMMs, we can obtain a physically more meaningful assessment of their acoustic similarity.

Adopting KLD for defining dissimilarity, the corresponding training criterion is referred as Minimum Divergence (MD) [Du *et al*. 2006; Du *et al*. 2007]. The criterion possesses the following potential advantages: 1) It employs acoustic similarity for high-resolution error definition, which is directly related to acoustic model refinement; 2) Label comparison is no longer used, which alleviates the influence of the chosen language model and phone set and the resultant hard binary decisions caused by label matching. Due to these advantages, MD is expected to be more flexible and robust.

In our work, MWE, which matches the evaluation metric, and MD, which focuses on refining acoustic dissimilarity, are compared. Other issues related to robust discriminative training, including how to design the maximum likelihood baseline and how to treat with the silence model is also discussed.

Experiments were performed on Aurora2 [Hirsch *et al*. 2000], which is a widely adopted database for research on noise robustness. For completeness, we tested the effectiveness of discriminative training on different ML baselines and different noise environments.

The rest of paper is organized as follows. In Section 2, issues on noise robustness of minimum error training will be discussed. In Section 3, MD training will be introduced. Experimental results are shown and discussed in Section 4. Finally, in Section 5, we give our conclusions.

2. Noise Robustness Analysis of Minimum Error Training

In this section, we will give a brief discussion of the major issues we are facing in robust discriminative training.

2.1 Error Resolution of Minimum Error Training

In [Povey 2004] and [Du *et al*. 2006], various discriminative training approaches are unified under the framework of minimum error training, where the objective function is an average of the recognition accuracies $\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_\mathrm{r})$ of all hypotheses weighted by the posterior probabilities. For conciseness, we consider the single training utterance case:

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{\boldsymbol{W}\in\mathcal{M}} P_{\boldsymbol{\theta}}(\boldsymbol{W}\mid\boldsymbol{O})\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_\mathrm{r}) \tag{1}$$

where $\boldsymbol{\theta}$ represents the set of the model parameters; $\boldsymbol{O}$ is a sequence of acoustic observation vectors; $\boldsymbol{W}_\mathrm{r}$ is the reference word sequence; $\mathcal{M}$ is the hypotheses space; $P_{\boldsymbol{\theta}}(\boldsymbol{W}\mid\boldsymbol{O})$ is the posterior probability of the hypothesis $\boldsymbol{W}$ given $\boldsymbol{O}$, which can be formulated as:

$$P_{\boldsymbol{\theta}}(\boldsymbol{W}\mid\boldsymbol{O}) = \frac{P_{\boldsymbol{\theta}}^{\kappa}(\boldsymbol{O}\mid\boldsymbol{W})P(\boldsymbol{W})}{\sum_{\boldsymbol{W}'\in\mathcal{M}} P_{\boldsymbol{\theta}}^{\kappa}(\boldsymbol{O}\mid\boldsymbol{W}')P(\boldsymbol{W}')} \tag{2}$$

where κ is the acoustic scaling factor.

The gain function $\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_r)$ is an *accuracy* measure of $\boldsymbol{W}$ given its reference $\boldsymbol{W}_r$. In Table 1, comparison of several minimum error criteria are listed. In MWE training, $\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_r)$ is word accuracy, which matches the commonly used evaluation metric of speech recognition. However, MPE has been shown to be more effective in reducing recognition errors because it provides a more precise measurement of word errors at the phone level. We can argue this point by advocating the final goal of discriminative training. In refining acoustic models to obtain better performance, it makes more sense to measure acoustic similarity between hypotheses instead of word accuracy. The symbol matching does not relate acoustic similarity with recognition. The measured errors can also be strongly affected by the phone set definition and language model selection. Therefore, acoustic similarity is proposed as a finer and more direct error definition in MD training.

Table 1. Comparison of criteria of minimum error training. (P_W : Phone sequence corresponding to word sequence W; LEV(,): Levenshtein distance between two symbol strings;| · |: Number of symbols in a string.)

Criterion	$\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_r)$	Objective
String based MCE	$\delta(\boldsymbol{W}=\boldsymbol{W}_r)$	Sentence accuracy
MWE	$\lvert\boldsymbol{W}_r\rvert-\mathrm{LEV}(\boldsymbol{W},\boldsymbol{W}_\mathrm{r})$	Word accuracy
MPE	$\lvert P_{\boldsymbol{W}_\mathrm{r}}\rvert-\mathrm{LEV}(P_{\boldsymbol{W}},P_{\boldsymbol{W}_\mathrm{r}})$	Phone accuracy
MD	$-D(\boldsymbol{W}_\mathrm{r}\parallel\boldsymbol{W})$	Acoustic similarity

Here, we aim at seeking how criteria with different error resolution performs in noisy environments. In our experiments, the whole-word model, which is commonly used in digit tasks, is adopted. For the noisy robustness analysis, MWE, which matches with the evaluation metric of speech recognition, will compared with MD, which possesses the highest error resolution as shown in Table 1.

2.2 Training Modes

In noisy environments, various ML trained baselines can be designed. So, the effectiveness of minimum error training with different training modes will be explored. In [Hirsch *et al.* 2000], two different sets of training, clean-training and multi-training, are used. In clean-training mode, only clean speech is used for training. Hence, there will be a mismatch when the model is tested in noisy environments. To alleviate the mismatch, multi-training, in which training set is composed of noisy speech with different SNRs, can be applied. Actually, multi-training can only achieve a "global SNR" match. To achieve a "local SNR" match, we adopt a SNR-based training mode. In the training phase, we train a series of models at different SNR levels, while in testing, all these models are paralleled as multi pronunciations of a HMM. Ideally, the model that matched the local SNR best will be automatically selected in decoding. SNR-based training can be considered as a high resolution acoustic modeling of multi-training. An illustration of the three training modes is shown in Figure 1.

An important issue in discriminative training is how to update silence or background models, which is even more critical in a noisy environment. In our research, we pay special attention to this issue for appropriate guidelines.

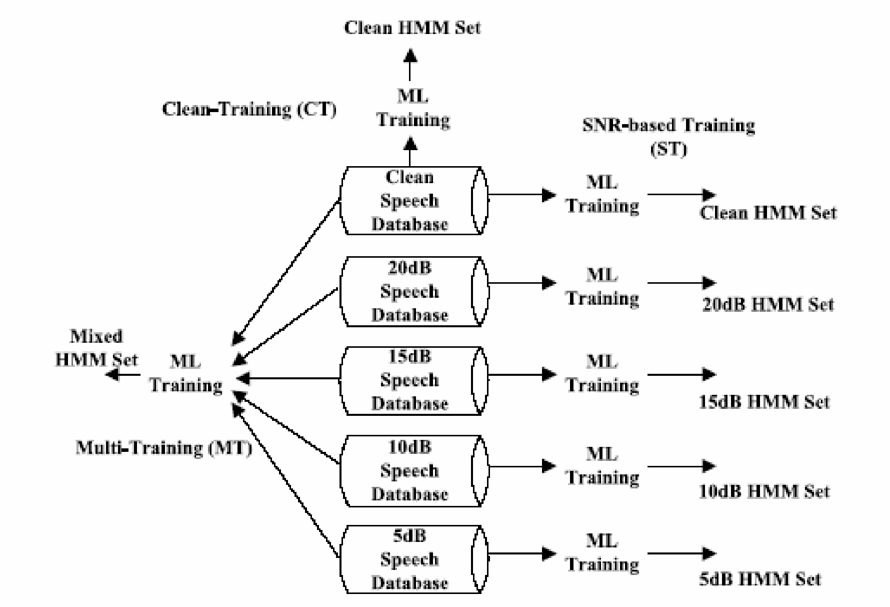


Figure 1. Illustration of three training modes

3. Word Graph based Minimum Divergence Training

3.1 Defining Errors by Acoustic Similarity

A word sequence is acoustically characterized by a sequence of HMMs. For automatically measuring acoustic similarity between $\boldsymbol{W}$ and $\boldsymbol{W}_{\mathrm{r}}$, we adopt KLD between the corresponding HMMs:

$$\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_{\mathrm{r}}) = -D(\boldsymbol{W}_{\mathrm{r}} \,||\, \boldsymbol{W}) \tag{3}$$

The HMMs, when they are reasonably well trained in ML sense, can serve as succinct descriptions of data.

3.2 KLD between Two Word Sequences

Our goal is to measure the KLD for word sequences in Eq. 3. Given two word sequences $\boldsymbol{W}_{\mathrm{r}}$ and $\boldsymbol{W}$ without their state segmentations, we should use a state matching algorithm to measure the KLD between the corresponding HMMs [Liu *et al.* 2005]. With state segmentations, the calculation can be further decomposed down to the state level:

$$\begin{aligned} D(\boldsymbol{W}_{\mathrm{r}} \,||\, \boldsymbol{W}) &= D(\boldsymbol{s}_{\mathrm{r}}^{1:T} \,||\, \boldsymbol{s}^{1:T}) \\ &= \int p(\boldsymbol{o}^{1:T} \mid \boldsymbol{s}_{\mathrm{r}}^{1:T}) \log \frac{p(\boldsymbol{o}^{1:T}|\boldsymbol{s}_{\mathrm{r}}^{1:T})}{p(\boldsymbol{o}^{1:T}|\boldsymbol{s}^{1:T})} d\boldsymbol{o}^{1:T} \end{aligned} \tag{4}$$

where T is the number of frames; $\boldsymbol{o}^{1:T}$ and $\boldsymbol{s}_{\mathrm{r}}^{1:T}$ are the observation sequence and hidden state sequence, respectively.

By assuming all observations are independent, we obtain:

$$D(\boldsymbol{s}_{\mathrm{r}}^{1:T} \,||\, \boldsymbol{s}^{1:T}) = \sum_{t=1}^{T} D(s_{\mathrm{r}}^{t} \,||\, s^{t}) = \sum_{t=1}^{T} \int p(\boldsymbol{o}^{t} \mid \boldsymbol{s}_{\mathrm{r}}^{t}) \log \frac{p(\boldsymbol{o}^{t} \mid \boldsymbol{s}_{\mathrm{r}}^{t})}{p(\boldsymbol{o}^{t} \mid \boldsymbol{s}^{t})} d\boldsymbol{o}^{t} \tag{5}$$

which means we can calculate KLD state by state, and sum them up.

Now, our problem is how to measure the KLD between two states. Conventionally, each state s is characterized by a Gaussian Mixture Model (GMM): $p(\boldsymbol{o} \mid s) = \sum_{m=1}^{M_s} w_{sm} \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\boldsymbol{sm}}, \boldsymbol{\Sigma}_{\boldsymbol{sm}})$, so the comparison is reduced to measuring KLD between two GMMs. Since there is no closed-form solution, we need to resort to the computationally intensive Monte-Carlo simulations. The unscented transform mechanism [Goldberger *et al.* 2003] has been proposed to approximate the KLD measurement of the two GMMs.

Let $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a N-dimensional Gaussian distribution and h be an arbitrary $\mathrm{IR}^{N} \rightarrow \mathrm{IR}$ function, the unscented transform mechanism suggests approximating the expectation of h by:

$$\int \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) h(\boldsymbol{o}) d\boldsymbol{o} \approx \frac{1}{2N} \sum_{k=1}^{2N} h(\boldsymbol{o}_k) \tag{6}$$

where $\boldsymbol{o}_k \, (1 \le k \le 2N)$ are the artificially chosen "*sigma*" points: $\boldsymbol{o}_k = \boldsymbol{\mu} + \sqrt{N\lambda_k}\, \boldsymbol{u}_k$, $\boldsymbol{o}_{k+N} = \boldsymbol{\mu} - \sqrt{N\lambda_k}\, \boldsymbol{u}_k \, (1 \le k \le N)$, where $\lambda_k, \boldsymbol{u}_k$ are the k^{th} eigenvalue and eigenvector of $\boldsymbol{\Sigma}$, respectively. Geometrically, all these "*sigma*" points are on the principal axes of $\boldsymbol{\Sigma}$. Equation 6 is precise if h is quadratic.

For our case, the Gaussian distribution in Eq. 6 is replaced by a GMM, and the function h corresponds to the term $\log \frac{p(\boldsymbol{o}^t | \boldsymbol{s}_{\mathrm{r}}^t)}{p(\boldsymbol{o}^t | \boldsymbol{s}^t)}$ in Eq. 5. Then, KLD between two states (GMMs) can be approximated by:

$$D(s_{\mathrm{r}}^{t} \,||\, s^{t}) \approx \frac{1}{2N} \sum_{m=1}^{M_{s_{\mathrm{r}}^{t}}} w_{s_{\mathrm{r}}^{t} m} \sum_{k=1}^{2N} \log \frac{p(\boldsymbol{o}_{mk} | s_{\mathrm{r}}^{t})}{p(\boldsymbol{o}_{mk} | s^{t})} \tag{7}$$

where $\boldsymbol{o}_{mk}$ is the k^{th} "*sigma*" point in the m^{th} Gaussian kernel of state s_{r}^{t}. By plugging this into Eq. 4, we obtain the KLD between two word sequences given their state segmentations.

3.3 Gain Function Calculation

Usually, a word graph is a compact representation of large hypotheses space in speech recognition. As the KLD between a hypothesised word sequence and the reference can be decomposed down to the frame level, we have the following word graph based representation of (1):

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{w \in \mathcal{M}} \sum_{\boldsymbol{W} \in \mathcal{M}: w \in \boldsymbol{W}} P_{\boldsymbol{\theta}}(\boldsymbol{W} \mid \boldsymbol{O}) \mathcal{A}(w) \tag{8}$$

where $\mathcal{A}(w)$ is the gain function of word arc w. Denoting b_w, e_w, the start frame index and end frame index of w, we have:

$$\mathcal{A}(w) = -\sum_{t=b_w}^{e_w} D(s_w^t \parallel s_{\mathrm{r}}^t) \tag{9}$$

where the s_w^t and s_{r}^t represent the certain state at time t on arc w and the reference, respectively.

From the objective function defined in Eq. 1, the gain function $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$ is dependent on the model parameters, which should be updated in optimization process. In [Du *et al.* 2007], we conclude that the optimization of the gain function term has little impact on the performance. So here, $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$ is considered a constant term and not optimized. The KLDs related to gain function are precomputed using the ML trained model parameters. Then our optimization of objective function is the same as that mentioned in [Povey 2004]. We use the Forward-Backward algorithm to update the word graph and the Extended Baum-Welch algorithm to update the model parameters in the training iterations.

4. Experiments

4.1 Experimental Setup

Experiments on TIDigits and Aurora2, both English continuous digit tasks, were performed. The English vocabulary is made of the 11 digits, from 'one(1)' to 'nine(9)', plus 'oh(0)' and 'zero(0)'. The baseline configuration for two databases is listed in Table 2.

Table 2. Baseline configuration

System	Feature	Model Type	# State /Digit	# Gauss /State	# string of training set	# string of testing set
TIDigits	MFCC_E_D_A	left-to-right whole-word model	10	6	12549	12547
Aurora2			16	3	8440*2	1001*70

The Aurora2 task consists of English digits in the presence of additive noise and linear convolutional channel distortion. These distortions have been synthetically introduced to clean

TIDigits data. Three testing sets measure performance against noise types similar to those seen in the training data (set A), different from those seen in the training data (set B), and with an additional convolutional channel (set C). The baseline performance and other details can be found in [Hirsch *et al*. 2000].

For minimum error training, the acoustic scaling factor κ was set to $\frac{1}{33}$. All KLDs between any two states were precomputed to make the MD training more efficient. For Aurora2, we select the best results after 20 iterations for each sub set of testing.

4.2 Experiments on TIDigits Database

As a preliminary result of noise robustness analysis, we first give the results of MD on the clean TIDigits database compared with MWE. As shown in Figure 2, performance of MD achieves 57.8% relative error reduction compared with the ML baseline and also outperforms MWE in all iterations.

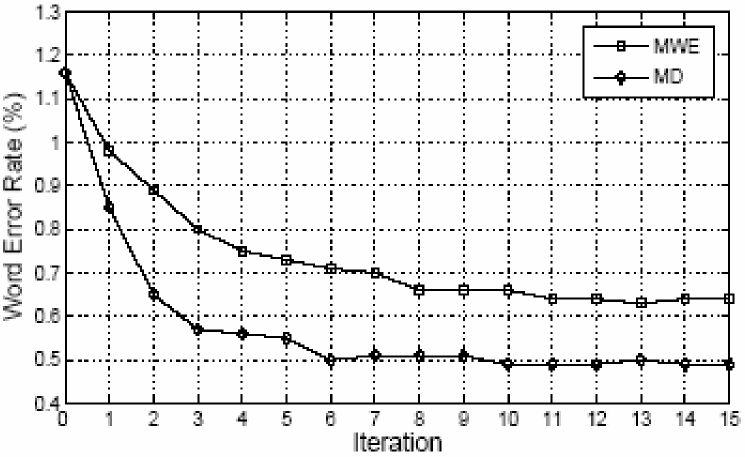


Figure 2. Performance comparison on TIDigits

4.3 Experiments on Aurora2 Database

Table 3. Word Accuracy (%) of MWE with or without silence model update in different training modes on Aurora2.

Training Mode	Update Silence Model	Set A	Set B	Set C	Overall
Clean	YES	61.85	56.94	66.26	60.77
Clean	NO	64.74	61.69	67.95	64.16
Multi	YES	89.15	89.16	84.66	88.26
Multi	NO	88.91	88.55	84.43	87.87

Silence Model Update. As shown in Table 3, we explore whether to update the silence model in minimum error training using different training modes. Since it is unrelated to the criteria, here we adopt MWE. When applying clean-training, the performances of all test sets without updating silence model are consistently better. However, in multi-training, the conclusion is the opposite. From the results, we can conclude that increasing the discrimination of the silence model will lead to performance degradation in mismatched cases (clean-training) and performance improvement in matched cases (multi-training). This can be explained as follows: For the clean-training case, if we increase the discrimination of the silence model, the noise segments are more easily recognized as digits when testing on noisy data. Then, insertion errors will increase. However, for the multi-training case, the silence model represents both silence and noise segments, which is matched with that when testing on noisy data. So, by updating the silence model, the global performance will be improved. Obviously, our SNR-based training belongs to the latter. In all our experiments, the treatment of silence model will obey this conclusion.

Table 4. Performance comparison on Aurora2 (MD vs. MWE)

Multi-Training – Results (Minimum Divergence)															
	A					B					C				Rel
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Impr
Clean	99.14	99.12	98.9	99.2	99.09	99.14	99.12	98.9	99.2	99.09	98.89	98.85	98.87	99.05	35.32%
20dB	98.71	98.55	98.81	98.61	98.67	98.43	98.37	98.57	98.89	98.57	98.65	97.64	98.15	98.52	43.92%
15dB	98.5	98	98.33	97.93	98.19	98	97.76	97.79	97.93	97.87	97.88	96.74	97.31	97.89	42.04%
10dB	97.18	96.55	97.2	96.08	96.75	96.41	95.8	96.06	95.31	95.90	95.15	94.04	94.60	95.98	34.81%
5dB	92.39	89.81	90.49	90.25	90.74	89.28	87.06	90.52	87.23	88.52	84.68	82.56	83.62	88.43	20.78%
0dB	72.8	64.63	58.93	70.32	66.67	65.24	64	69.19	62.48	65.23	49.25	54.44	51.85	63.13	10.51%
-5dB	31.04	29.56	22.7	28.57	27.97	30.06	28.96	33.58	25.46	29.52	22.01	24.24	23.13	27.62	4.15%
Average	91.92	89.51	88.75	90.64	90.20	89.47	88.60	90.43	88.37	89.22	85.12	85.08	85.10	88.79	
Rel Impr	28.10%	12.93%	16.53%	21.79%	19.60%	27.93%	12.04%	22.53%	22.40%	21.45%	11.21%	4.93%	8.17%		17.62%

Multi-Training – Results (Minimum Word Error)															
	A					B					C				Rel
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Impr
Clean	99.14	99.18	99.02	99.29	99.16	99.14	99.18	99.02	99.29	99.16	98.99	99.06	99.03	99.13	40.96%
20dB	98.86	98.67	98.78	98.7	98.75	98.74	98.43	98.72	98.95	98.71	98.34	97.4	97.87	98.56	45.45%
15dB	98.74	98.13	98.33	97.69	98.22	98.5	97.82	98.03	98.06	98.10	97.33	96.25	96.79	97.89	41.97%
10dB	96.87	95.95	96.87	95.43	96.28	96.22	95.53	96.42	95.74	95.98	94.63	93.5	94.07	95.72	30.03%
5dB	92.32	88.85	88.25	88.83	89.56	88.36	87.3	89.53	86.61	87.95	84.49	82.62	83.56	87.72	15.40%
0dB	70.31	63.33	53.44	64.7	62.95	64.6	68.18	68.27	59.12	65.04	47.62	54.44	51.03	61.40	6.25%
-5dB	29.66	29.72	21.8	25.27	26.61	30.21	27.84	33.49	23.97	28.88	21.31	24.24	22.78	26.75	3.01%
Average	91.42	88.99	87.13	89.07	89.15	89.28	89.45	90.19	87.70	89.16	84.48	84.84	84.66	88.26	
Rel Impr	23.69%	8.60%	4.53%	8.69%	10.98%	26.64%	18.62%	20.65%	17.92%	21.02%	7.39%	3.39%	5.46%		13.71%

Error Resolution of Minimum Error Training. As shown in Table 4, the performances of MD and MWE are compared. Here, multi-training is adopted because it is believed that matching between training and testing can tap the potential of minimum error training. For the overall performance on three test sets, MD consistently outperforms MWE. From the viewpoint of SNRs, MD outperforms MWE in most cases when SNR is below 15dB. Hence, we can conclude that, although MWE matches with the model type and evaluation metric of speech recognition, MD, which possesses the highest error resolution, outperforms it in low SNR. In other words, the performance can be improved in low SNR by increasing the error resolution of criterion in minimum error training. This conclusion can be also drawn in clean-training and SNR-based training cases.

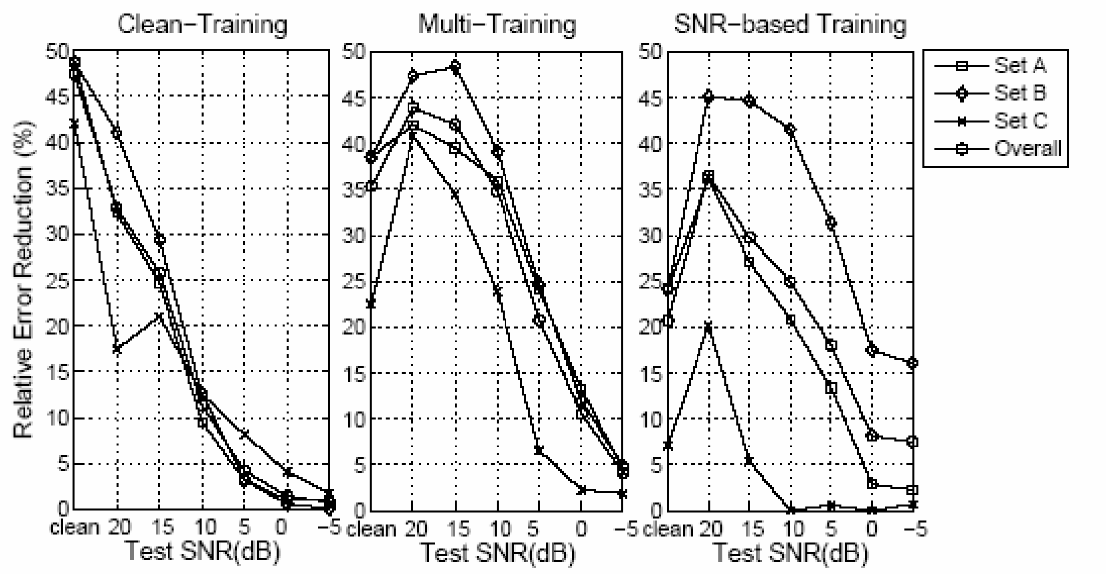


Figure 3. Relative Improvement over ML baseline on Aurora2 using different training modes in MD training

Table 5. Summary of performance on Aurora2 using different training modes in MD training.

	Word Accuracy (%)	Relative Improvement
Training Mode	Set A Set B Set C Overall	Set A Set B Set C Overall
Clean-Training	63.49 58.94 68.96 62.76	5.56% 7.21% 8.32% 6.76%
Multi-Training	90.20 89.22 85.10 88.79	19.60% 21.45% 8.17% 17.62%
SNR-based Training	91.27 89.27 86.70 89.56	10.00% 26.21% 1.14% 15.68%

Different Training Modes. Figure 3 shows relative improvement over ML baseline using MD training with different training modes. From this figure, some conclusions can be obtained. First, set B, whose noise scenarios are different from training, achieves the most obvious relative improvement in most cases. The relative improvement of set A is comparable with set B in the clean-training and multi-training, but worse than set B in SNR-based training. The relative improvement of set C, due to the mismatch of noise scenario and channel, was almost the worst in all training modes. Second, the relative improvement performance declines for decreasing SNR in clean-training. However, in multi-training and SNR-based training, the peak performance is in the range of 20dB to 15dB. Also, in the low SNRs, the performance of cleaning-training is worse than the other two training modes on set A and set B.

The summary of performance is listed in Table 5. Word accuracy of our SNR-based training outperforms multi-training on all test sets, especially set A and set C. For the overall relative improvement, the best result of 17.62% is achieved in multi-training.

5. Conclusions

In this paper, the noise robustness of discriminatively trained HMMs is investigated. Discriminatively trained models are tested on English continuous digit databases, and MD and MWE criteria are experimentally compared to test the affection of error resolution. We observe: 1. Minimum error training is effective not only in clean environments, but also in noisy environments, which can be concluded in various training modes. Minimum error training is more effective as the SNR increases. Even when testing on mismatched noise scenarios, minimum error training also achieves better performance than ML training. 2. In minimum error training, higher resolution error analysis is more helpful at low SNRs. 3. Silence models should be carefully updated when the training and testing data are not well-matched.

Reference

Du, J., P. Liu, H. Jiang, F.K. Soong, and R.-H. Wang, "A New Minimum Divergence Approach to Discriminative Training," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2007, pp. 677-680.

Du, J., P. Liu, F.K. Soong, J.-L. Zhou, and R.-H. Wang, "Minimum Divergence Based Discriminative Training," In*Proceedings of International Conference on Spoken Language Processing*, 2006, pp. 2410-2413.

Gales, M.J.F. and S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," Technical Report EDICS Number: SA 1.6.8, Cambridge University, 1994.

Goldberger, J., "An Efficient Image Similarity Measure Based on Approximations of KL-Divergence between Two Gaussian Mixtures," In*Proceedings of International Conference on Computer Vision*, 2003, pp. 370-377.

Gong, Y., "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, 16, 1995, pp. 261-291.

Hirsch, H.G. and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," In*Proceedings of ISCA ITRW ASR*, 2000, pp. 181-188.

Juang, B.-H., W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recogtion," *IEEE Transactions on Speech and Audio Processing*, 5(3), 1997, pp. 257-265.

Kullback, S. and R.A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat*, 22, 1951, pp. 79-86.

Laurila, K., M. Vasilache, and O. Viikki, "A Combination of Discriminative and Maximum Likelihood Techniques for Noise Robust Speech Recognition," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1998, pp. 85-88.

Liu, P., F.K. Soong, and J.-L. Zhou, "Effective Estimation of Kullback-Leibler Divergence between Speech Models," Technical Report, Microsoft Research Asia, 2005.

Meyer, C. and G. Rose, "Improved Noise Robustness by Corrective and Rival Training," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2001, pp. 293-296.

Ohkura, K., D. Rainton, and M. Sugiyama, "Noise-robust HMMs Based on Minimum Error Classification," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1993, pp. 75-78.

Povey, D., "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University, 2004.

Schluter, R., "Investigations on Discriminative Training Criteria," PhD thesis, Aachen University, 2000.

Valtchev, V., J.J. Odell, P.C. Woodland, and S.J. Young, "MMIE Training of Large Vocabulary Speech Recognition Systems," *Speech Communication*, 22, 1997, pp. 303-314.

Varga, A.P. and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1990, pp. 845-848.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 303-324

Multilingual Spoken Language Corpus Development for Communication Research

Toshiyuki Takezawa[*], **Genichiro Kikui**[*+], **Masahide Mizushima**[*+], and **Eiichiro Sumita**[#*]

Abstract

Multilingual spoken language corpora are indispensable for research on areas of spoken language communication, such as speech-to-speech translation. The speech and natural language processing essential to multilingual spoken language research requires unified structure and annotation, such as tagging. In this study, we describe an experience with multilingual spoken language corpus development at our research institution, focusing in particular on speech recognition and natural language processing for speech translation of travel conversations. An integrated speech and language database, Spoken Language DataBase (SLDB) was planned and constructed. Basic Travel Expression Corpus (BTEC) was planned and constructed to cover a variety of situations and expressions. BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their translations, and SLDB is a collection of transcriptions of bilingual spoken dialogs. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, *i.e.*, hotel situations. BTEC contains approximately 588k utterance-style expressions, while SLDB contains about 16k utterances. Machine-aided Dialogs (MAD) was developed as a development corpus, and both BTEC and SLDB can be used to handle MAD-type tasks. Field Experiment Data (FED) was developed as the evaluation corpus. We conducted an experiment, and based on analysis of our follow-up questionnaire, roughly half the subjects of the

[*] ATR Spoken Language Communication Research Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan Telephone: +81 774 95 1301 Fax: +81 774 95 1308
E-mail: toshiyuki.takezawa@atr.jp

[+] currently with NTT Cyberspace Laboratories, Japan
E-mail: {kikui.genichiro, mizushima.masahide}@lab.ntt.co.jp

[#] National Institute of Information and Communications Technology, Japan
E-mail: eiichiro.sumita@{nict.go.jp; atr.jp}

[Received April 6, 2007; Revised August 3, 2007; Accepted August 6, 2007]

experiment felt they could understand and make themselves understood by their partners.

Keywords: Multilingual Corpus, Spoken Language, Speech Translation, Dialog, Communication.

1. Introduction

Various kinds of corpora developed for analysis of linguistic phenomena and statistical information gathering are now accessible via electronic media and can be utilized for the study of natural language processing. Since these include written-language and monolingual corpora, however, they are not necessarily useful for research and development of multilingual spoken language processing. A multilingual spoken language corpus is indispensable for research on areas of spoken language communication such as speech-to-speech translation.

Research on speech translation began in the 1980s. NEC demonstrated a prototype speech translation system at the Telecom ’83 exhibition. ATR Interpreting Telephony Research Laboratories was established in 1986 for the research of basic speech translation technologies and produced ASURA [Morimoto *et al.* 1993]. This system can recognize well-formed Japanese utterances in a limited domain, translate them into both English and German, and output synthesized speech. The ASURA system was used for the International Joint Experiment of Interpreting Telephony with participants from Kyoto, Japan (ATR), Pittsburgh, USA (Carnegie Mellon University [Lavie *et al.* 1997]) and Munich, Germany (Siemens and University of Karlsruhe) in January 1993 [Morimoto *et al.* 1993].

Many projects on speech-to-speech translation began at that time [Rayner *et al.* 1993; Roe *et al.* 1992; Wahlster *et al.* 2000]. SRI International and Swedish Telecom developed a prototype speech translation system that could translate queries from spoken English to spoken Swedish in the domain of air travel information systems [Rayner *et al.* 1993]. AT&T Bell Laboratories and Telefónica Investigación y Desarrollo developed a restricted domain spoken language translation system called Voice English/Spanish Translator (VEST) [Roe *et al.* 1992]. In Germany, *Verbmobil* [Wahlster 2000], was created as a major speech-to-speech translation research project. The *Verbmobil* scenario assumes native speakers of German and of Japanese who both possess at least a basic knowledge of English. The *Verbmobil* system supports them by translating from their mother tongue, *i.e.* Japanese or German, into English.

In the 1990s, speech recognition and synthesis research shifted from a rule-based to a corpus-based approach such as HMM and N-gram. However, machine translation research still depended mainly on a rule-based or knowledge-based approach. In the 2000s, wholly corpus-based projects such as European TC-STAR [Höge 2002; Lazzari 2006] and DARPA GALE [Roukos 2006] began to deal with monologue speeches such as broadcast news and

European Parliament plenary speeches. In this paper, we report corpus construction activities for translation of spoken dialogs of travel conversations.

There are a variety of requirements for every component technology, such as speech recognition and language processing. A variety of speakers and pronunciations may be important for speech recognition, and a variety of expressions and information on parts of speech may be important for natural language processing. The speech and natural language processing essential to multilingual spoken language research requires unified structure and annotation, such as tagging.

In this paper, we introduce an interpreter-aided spoken dialog corpus and discuss corpus configuration. Next, we introduce the basic travel expression corpus developed to train machine translation of spoken language among Japanese, English, and Chinese speakers. Finally, we discuss the Japanese, English, and Chinese multilingual spoken dialog corpus that we created using speech-to-speech translation systems.

2. Overview of Approach

We first planned and constructed an integrated speech and language database called Spoken Language DataBase (SLDB) [Morimoto *et al.* 1994; Takezawa *et al.* 1998]. The task involved travel conversations between a foreign tourist and a front desk clerk at a hotel; this task was selected because people are familiar with it and because we expect it to be included in future speech translation systems. All of the conversations for this database take place in English and Japanese through interpreters because the research at that time concentrated on Japanese and English. The interpreters serve as the speech translation system. One remarkable characteristic of the database is its integration of speech and linguistic data. Each conversation includes data on recorded speech, transcribed utterances, and their correspondences. This kind of data is very useful because it contains transcriptions of spoken dialogs between speakers who speak different mother tongues. However, the cost of collecting spoken languages is too high to expand the size.

There are three important points to consider in designing and constructing a corpus for dialog-style speech communication such as speech-to-speech translation. The first is to have a variety of speech samples with a wide range of pronunciations, speaking styles, and speakers. The second point is to have data for a variety of situations. A “situation” means one of various limited circumstances in which the system’s user finds him- or herself, such as an airport, a hotel, a restaurant, a shop, or in transit during travel; it also involves various speakers’ roles, such as communication with a middle-aged stranger, a stranger wearing jeans, a waiter or waitress, or a hotel clerk. The third point is to have a variety of expressions.

According to our previous study [Takezawa *et al.* 2000], human-to-machine conversational speech data shared characteristics with human-to-human indirect communication speech data such as spoken dialogs between Japanese and English speakers through human interpreters. Moreover, human-to-human indirect communication data had an intermediate characteristic, *i.e.*, it was positioned somewhere between direct communication data, that is, Japanese monolingual conversations, and speech data from conversational text. If we assume that a speaker would accept a machine-friendly speaking style, we could take a great step forward: a clear separation of speech data collection and multilingual data collection. In the following, we focus on multilingual data collection. In order, Basic Travel Expression Corpus (BTEC) [Takezawa *et al.* 2002; Kikui *et al.* 2003] was planned to cover the varieties of situations and expressions.

Machine-aided Dialogs (MAD) was planned as a development corpus to handle the differences between the target utterance with which speech translation systems must deal and the following two corpora.

SLDB contains no recognition/translation errors because the translations between people speaking different languages are done by professional human interpreters. However, even a state-of-the-art speech translation system cannot avoid recognition/translation errors.

BTEC contains edited colloquial travel expressions, which are not transcriptions, so some people might not express things in the same way, and the frequency distribution of expressions might be different from actual dialogs.

Field Experiment Data (FED) was planned as the evaluation corpus. Table 1 shows an overview of the corpora. In the table, S2ST stands for speech-to-speech translation, MT stands for machine translation, J, E, and C stand for Japanese, English, and Chinese, respectively.

Table 1. Overview of corpora

Name	SLDB	BTEC	MAD	FED
	Spoken Language DataBase	Basic Travel Expression Corpus	Machine-Aided Dialogs	Field Experiment Data
Purpose	Developing S2ST	Training MT	Developing S2ST	Evaluation of S2ST
Domain	Hotel	Travel	Travel	Travel
Languages	J E (C)	J E C	J E (C)	J E C
Speaker Participants	71 (+23 Interpreters)	Not spoken	45	84
Size	16k	588k	13k	2k

3. Interpreter-Aided Spoken Dialog Corpus (SLDB)

SLDB contains data from dialog spoken between English and Japanese speakers through human interpreters [Morimoto *et al.* 1994; Takezawa *et al.* 1998]. All utterances in SLDB have been translated into Chinese. The content is entirely travel conversations between a foreign tourist and a front desk clerk at a hotel. Human interpreters serve as the speech translation system.

Table 2 is an overview of the corpus, and Table 3 shows its basic characteristics.

Table 2. Overview of SLDB

Number of collected dialogs	618
Speaker participants	71
Interpreter participants	23

Table 3. Basic characteristics of SLDB

	Japanese	English
Number of utterances	16,084	16,084
Number of sentences	21,769	22,928
Number of word tokens	236,066	181,263
Number of word types	5,298	4,320
Average number of words per sentence	10.84	7.91

One remarkable characteristic of SLDB is its integration of speech and linguistic data. Each conversation includes recorded speech data, transcribed utterances, and the correspondences between them.

The transcribed Japanese and English utterances are tagged with morphological information. This kind of tagged information is crucial for natural language processing as well as for speech recognition language modeling. The recorded speech signals and transcribed utterances in the database provide both examples of various phenomena in bilingual conversations, and input data for speech recognition and machine translation evaluation purposes.

Data can be classified into the following three major categories.

1. Transcribed data
2. Tagged data
3. Speech data

Transcribed data consists of the following.

(a) Bilingual text

(b) Japanese text

(c) English text

The recorded bilingual conversations are transcribed into a text file. The bilingual text contains descriptions of the situations in which a speech translation system is used.

J: Arigatou gozaimasu. Kyoto Kankou Hotel de gozaimasu.

JE: Thank you for calling Kyoto Kanko Hotel. |How may I help you?

E: Good evening. |I'd like to make a reservation, please.

EJ: Konbanwa. |Yoyaku wo shi tai n desu keredomo.

J: Hai,[e-]go yoyaku no hou wa itsu desho u ka?

JE: Yes, when do you plan to stay?

E: I'd like to stay from August tenth through the twelfth, for two nights.| If possible, I'd like a single room, please.

EJ: Hachigatsu no tooka kara juuni-nichi made, ni-haku shi tai n desu.| Dekire ba, single room de onegaishimasu.

J: Kashikomarimashita. |Shoushou o-machi kudasai mase.

JE: All right, please wait a moment.

J: O-mata se itashimashita.| Osoreiri masu ga, single room wa manshitsu to nat te orimasu.

JE: I am very sorry our single rooms are all booked.

J: [e]Washitsu ka twin room no o-hitori sama shiyou deshi tara o-tori dekimasu ga.

JE: But, Japanese style rooms and twin rooms for single use are available.

E: [Oh] what are the rates on those types of rooms?

EJ: Sono o-heya no ryoukin wo oshie te kudasai.

Figure 1. Conversation between an American tourist and a Japanese front desk clerk.

Figure 1 shows an example of transcribed conversations. The Japanese text in Figure 1 has been transcribed into Romanized Japanese for the convenience of readers who do not understand Japanese *hiragana*, *katakana*, and *kanji* (Chinese characters). The original text was transcribed in Japanese characters *hiragana*, *katakana*, and *kanji*. Interjections are bracketed. J, E, JE, or EJ at the beginning of a line denotes a Japanese speaker, an English speaker, a Japanese-to-English interpreter, or an English-to-Japanese interpreter, respectively. "｜" denotes a sentence boundary. A blank line between utterances shows that the utterance's right was transferred.

The Japanese text is produced by extracting the utterances of a Japanese speaker and an English-to-Japanese interpreter, while the English text is produced by extracting the utterances of an English speaker and a Japanese-to-English interpreter. These two kinds of data are utilized for such monolingual investigations as morphological analysis.

The tagged data consists of the following.

(d) Japanese morphological data

(e) English morphological data

SLDB is available to outside research institutions and can be accessed at the following URL: http://www.atr.jp.

4. Basic Travel Expression Corpus (BTEC)

The Basic Travel Expression Corpus (BTEC) [Takezawa *et al.* 2002; Kikui *et al.* 2003] was designed to cover utterances for possible travel conversations topic and their translations. Since it is practically impossible to collect them by transcribing actual conversations or simulated dialogs, we decided to use sentences provided by bilingual travel experts based on their experience. We started by looking at phrasebooks that contain bilingual sentence pairs (in this case Japanese/English) that the editors consider useful for tourists traveling abroad. Such sentence pairs were collected and rewritten to make translation as context-independent as possible and to comply with the speech transcription style of our research institution. Sentences that were outside of the travel domain or have very special meanings were removed.

Table 4 lists the basic statistics of the BTEC collections, called BTEC1, 2, 3, 4, and 5. Each collection was created using the same procedure in a different time period or using a different translation direction from the source language to target languages. Strictly speaking, morphemes are used as the basic linguistic unit for Japanese (instead of words), since morpheme units are more stable than word units.

Table 4. Overview of BTEC

	BTEC1	BTEC2	BTEC3	BTEC4	BTEC5
Number of utterance-style expressions	172k	46k	198k	74k	98k
Number of Japanese word tokens	1,174k	341k	1,434k	548k	1,046k
Number of Japanese word types	28k	20k	43k	22k	28k
Languages (Source:Targets)	J:EC	J:EC	J:EC	E:JC	E:JC

The aims of the BTEC corpus are for translation and language modeling for automatic speech recognition. For translation, one of the key points to cover is the translation direction from the source language to target languages. For automatic speech recognition in the travel domain, one of the key points to cover is multiple sub-domains such as airport-related dialogs, hotel-related dialogs, and so on.

For translation, the BTEC collections cover both translation directions. BTEC1, BTEC2, and BTEC3 contain expressions for Japanese tourists visiting the USA, UK, or Australia. The translation direction is from Japanese to English and Chinese. BTEC4 mainly contains expressions for American tourists who visit Japan. The translation direction is from English to Japanese and Chinese. BTEC5 contains various expressions, such as those for American tourists who go to Korea. The translation direction is from English to Japanese and Chinese.

For automatic speech recognition, BTEC covers multiple domains. Domain information is given for BTEC1, BTEC2, and BTEC3. Table 5 shows an overview.

BTEC sentences, as described above, did not come from actual conversations but were generated by experts as reference materials. This approach enabled us to efficiently create a broad corpus; however, it may have two problems. First, this corpus may lack utterances that occur in real conversation. For example, when people ask the way to a bus stop, they often use a sentence like (1). However, in BTEC this is expressed more directly, as in (2).

(1) I'd like to go downtown. Where can I catch a bus?

(2) Where is a bus stop (to go downtown)?

We will discuss this issue in the section on MAD.

The second problem is that the frequency distribution of this corpus may be different from the actual distribution. In this corpus, the frequency of an utterance most likely reflects the best trade-off between usefulness in real situations and compactness of the collection. Therefore, it is possible to think of this frequency distribution as a first approximation of reality, but this is an open question.

A part of BTEC was distributed to the participants in the International Workshop on Spoken Language Translation (IWSLT) [IWSLT 2006].

Table 5. Domain information of BTEC

Domain	Frequency
Communication	20.4%
Basic	19.1%
Trouble	8.7%
Shopping	7.9%
Stay	6.9%
Sightseeing	6.6%
Transfer	6.6%
Restaurant	5.9%
Business	3.8%
Airport	3.6%
Contact	3.3%
Airplane	2.3%
Drink	1.0%
Home stay	1.0%
Exchange	0.8%
Snack	0.8%
Beauty	0.5%
Study overseas	0.5%
Go home	0.3%
Total	100.0%

5. Machine Translation-Aided Spoken Dialog Corpus (MAD)

The approach exemplified by BTEC focuses on maximizing the coverage of the corpus rather than creating an accurate sample of reality. Users may use different wording when they speak to the system. In addition, there may be differences between the target utterance with which speech translation systems must deal and the following two corpora.

SLDB contains no recognition/translation errors because the translations between people speaking different languages are done by professional human interpreters. However, even a state-of-the-art speech translation system cannot avoid recognition/translation errors.

BTEC contains edited colloquial travel expressions, which are not transcriptions, so some people might not express things in the same way and the frequency distribution of expressions might be different from actual dialogs.

Therefore, MAD is intended to collect representative utterances that people will input into S2ST systems. For this purpose, simulated dialogs (*i.e.*, role play) were carried out between two native speakers of different mother tongues with a Japanese/English bi-directional S2ST system, instead of using human interpreters.

During the first half of the research program, human typists were used instead of speech recognizers to ensure that we collected good quality data. During the second half of the research program, the S2ST system between English and Japanese was used.

5.1 Collecting Spoken Dialog Data Using Typists

Figure 2 is an overview of the data collection environment. An English typist transcribes an English utterance and inputs it into a machine translation system from English to Japanese. The translated Japanese text and its synthesized speech are sent to a Japanese speaker. Likewise, a Japanese typist transcribes a Japanese utterance and inputs it into a machine translation system from Japanese to English. The translated English text and its synthesized speech are sent to an English speaker. By repeating this process, an MT-aided bilingual dialog continues. Speech waves, transcriptions, and translated texts are stored in log files.

Five sets of simulated dialogs (MAD1 through MAD5) have so far been developed, changing parameters such as system configurations, complexity of dialog tasks, instructions to speakers, and so on. Table 6 shows a summary of the five experiments, MAD1-MAD5. In this table, the number of utterances includes both Japanese and English.

The first set of dialogs (MAD1) was collected to see whether conversation through a machine translation system is feasible. The second set (MAD2) focused on task achievement by assigning complex tasks to participants. The third set (MAD3) contains carefully recorded speech data of medium complexity. MAD4 and MAD5 aim to investigate how utterances change based on a change in setting.

It is very likely that people would speak differently to a spoken language system based on the instructions given to them. Instructions were conveyed to subjects for all sets other than MAD1 using instructional movies to ensure that the same instructions were given to each subject. Before starting the experiments, subjects were asked to watch these movies and then try the system with test dialogs. Instructions and practice took about 30 minutes. We gave different types of instructions for the fourth set (MAD4).

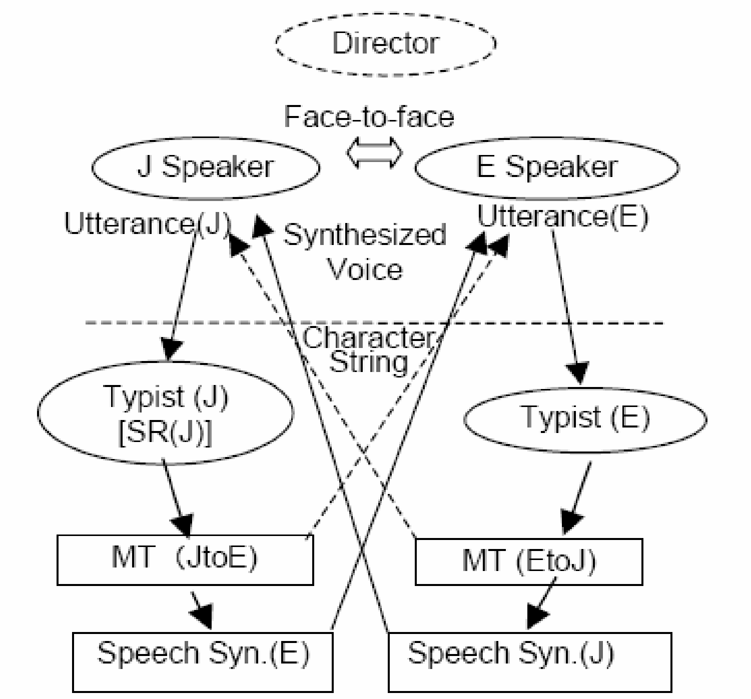


Figure 2. Data collection environment of MAD

Table 6. Statistics of MAD corpora

Subset ID	MAD1	MAD2	MAD3	MAD4	MAD5
Reference	[Takezawa and Kikui 2003]	[Takezawa and Kikui 2003]	[Takezawa *et al*. 2003]	[Takezawa and Kikui 2004]	[Mizushima *et al*. 2004]
Number of utterances	3,022	1,696	2,180	1,872	1,437
Number of words per utterance	10.0	12.6	11.1	9.82	8.47
Number of utterances per dialog	7.8	49.3	18.8	22.0	27.0
Task complexity	Simple	Complex	Medium	Medium	Medium

Average numbers depend on experimental conditions.

S2ST presupposes that each user understands the translated utterances of the other. However, the dialog environment described so far allows the user to access other information, such as translated text displayed on a PDA. We tried to control the extra information in MAD5 to see how utterances would be affected.

Part of the MAD corpus has been translated into Chinese.

5.2 Collecting Spoken Dialog Data Using Speech Translation Systems

Spoken dialog data was collected using the S2ST system for English and Japanese. This data collection experiment is called MAD6 because five data collection experiments were carried out using typists. The system was configured as follows.

- Acoustic model for Japanese speech recognition: Speaker-adapted models.
- Language model for Japanese speech recognition: Vocabulary size 52,000 words.
- Acoustic model for English speech recognition: Speaker-adapted models.
- Language model for English speech recognition: Vocabulary size 15,000 words.

Table 7 is an overview of MAD6. Data collected by typists (MAD1 through MAD5) contains some translation errors but very few recognition errors. However, MAD6 data contains both recognition errors and translation errors. We found that translation errors caused by recognition errors sometimes caused great confusion. That is, users need many more turns to recover from translation errors caused by recognition errors than to recover from mere translation errors. Moreover, we found that the user's speaking style changed similar to read speech when using speech recognizers. This was because users could confirm their recognition results using a PC display. Experienced users soon understood that they were confused by translation errors caused by recognition errors and adopted strategies to avoid recognition errors. As a result, their speaking style seemed to change from a natural dialog style to a read speech style.

Table 7. Overview of MAD6

	MAD6
Purpose	Spoken dialog data collection using S2ST system
Task	Simple as in MAD1
Number of utterances	2,507
Number of dialogs	139

5.3 Comparative Analysis and Discussion

BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their translations, and SLDB is a collection of transcriptions of bilingual spoken dialogs. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, *i.e.*, hotel situations. BTEC contains approximately 588k utterance-style expressions, while SLDB contains about 16k utterances. Thus, we can hypothesize that BTEC and SLDB together cover the same content as MAD. This hypothesis is partly validated by the cross-perplexity shown in Table 8. In this table, BTEC1 + SLDB combines two language models trained on BTEC1 and SLDB with linear interpolation. Similarly, BTEC1 + Extra combines BTEC1 and a corpus called Extra, which is a sample of a BTEC-type extra corpus of about the same size as SLDB. This clearly shows that both BTEC1 and SLDB are required for handling MAD-type tasks. Further discussion is available in [Kikui *et al*. 2006].

Table 8. Cross-perplexity for MAD (Japanese)

	Training corpus			
	BTEC1	SLDB	BTEC1 + SLDB	BTEC1 + Extra
Size (Number of utterances)	162k	12k	174k	174k
Cross-perplexity	38.2	94.9	30.7	35.7

6. Field Experiment Data (FED)

An ideal approach to applying a system to real utterances is to let people use the system in real world settings to achieve real conversational goals (*e.g.*, booking a package tour). This approach, however, has at least two problems. First, it is difficult to back up the system when it makes errors because current technology is not perfect. Second, it is difficult to control tasks and conditions to do meaningful analysis of the collected data.

The new experiment reported here was still in the role-play style but its dialog situations were designed to be more natural. The S2ST system for travel conversation was set up at tourist information centers in an airport and a train station, and non-Japanese-speaking people were asked to talk with the Japanese staff at information centers using the S2ST system.

6.1 Experimental System for Data Collection

Figure 3 is a diagram of the overall experimental system. The system includes two PDAs, one for each language, and several PC servers. The PC servers are controlled by a special controller called the gateway for component engines, consisting of automatic speech recognition (ASR) [Itoh *et al*. 2004], machine translation (MT) [Sumita *et al*. 2004], and speech synthesis (SS) [Kawai *et al*. 2004] PCs for each language and each language-pair. The gateway is responsible for controlling information flow between PDAs and engines. It is also responsible for mediating messages from the ASR and MT engines to PDAs. Each PDA is connected to the gateway with a wireless LAN. The gateway and component engines are wired. Headset microphones were used in the FED experiment.

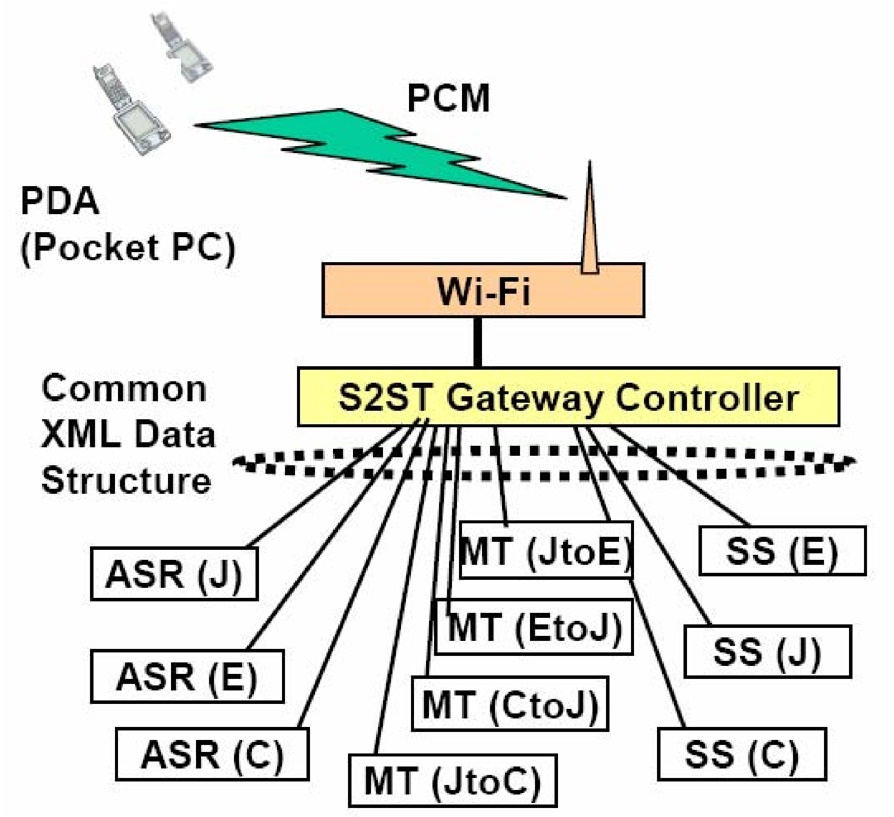


Figure 3. Overview of experimental system

An utterance spoken into a PDA is sent to the gateway server, which calls the ASR, MT, and SS engines in this order to have the utterance translated. Finally, the gateway sends the translated utterance to the other PDA.

Speaker-adapted acoustic models were used for Japanese speech recognition because only a few Japanese staff at the tourist office agreed to participate in the FED experiment. A few proper names that were deemed necessary to carry out the planned conversations were added to the lexicon. These included names such as those of stations near the locations of the experiment.

6.2 Locations

Data collection was conducted near two tourist information centers. One was in Kansai International Airport (hereafter, KIX), and the other was at Osaka City Air Terminal (hereafter, OCAT) in the center of Osaka. The former is in the main arrival lobby of the airport, which many tourists pass as they emerge from customs. The latter is a semi-enclosed area of about 40 m^2 enclosed by glass walls (but with two open doors).

Environmental noise was 60-65 dBA in both places but rose to 70 dBA when the public address system was in use.

The language pairs were English-Japanese/Japanese-English and Chinese-Japanese/Japanese-Chinese.

6.3 Scenario

A good method of collecting real utterances is to just let subjects talk freely without using predetermined scenarios. Analyzing uncontrolled dialog, however, is very difficult. In the FED experiment, eight dialog scenarios were prepared. These scenarios, listed below, are categorized by expected number of turns for each speaker into three levels of complexity.

Level-1 : Requires one or two turns per speaker plus greetings.
E.g., "Please ask where the bus stop for Kyoto station is."

Level-2 : Requires three or four turns per speaker plus greetings.
E.g., "Please ask the way to Kyoto station."

Level-3 : Free discussion.
E.g., "Please ask anything related to traveling in the Osaka area."

Real dialogs included many clarification sub-dialogs necessitated by incomprehensible output from the system. This means that the number of turns was actually larger than we expected or planned.

6.4 Speakers

6.4.1 Japanese Speakers

We asked staff at the tourist information centers to participate in the experiments, and six people at KIX and three at OCAT agreed to take part.

6.4.2 Chinese Speakers

Since the Chinese speech recognizer was trained on Mandarin speech, we needed to recruit subjects from the Beijing region of China. It was, however, difficult to find tourists from China who had time to participate in the experiment because most of them came to Osaka as members of tightly scheduled group tours. Therefore, we relied on 36 subjects gathered by the Osaka prefectural government. These subjects are college students from China majoring in non-technical areas such as foreign studies and tourism.

6.4.3 English Speakers

The English speech recognizer was trained on North American English. Again, however, it was difficult to find volunteer subjects who speak North American English. We expected to recruit many individual tourists, and most of the English-speaking volunteer subjects were indeed tourists arriving at or leaving the airport during the experiment. In addition to these volunteers, Osaka prefecture provided nine subjects who were working in Japan as English teachers. The resulting 39 subjects were not all North Americans, as shown in Table 9.

Table 9. Origin of English-speaking subjects

Origin	Number of subjects
USA	15
UK	6
Australia	5
Canada	4
New Zealand	2
Denmark	2
Other	5

6.5 Collecting Data

First, we set up the S2ST system and asked the Japanese subjects (*i.e.*, service personnel at the tourist information centers) to stand by at the experimental sites.

When an English or Chinese speaking subject visited a center, he or she was asked to fill out the registration form. Then, the staff explained for 2-3 minutes how to use the S2ST system and asked the subject to try very simple utterances like “hello” or “thank you.” After the trial utterances, we had the subject try two dialogs: one dialog for practice using a level 1 scenario, and the other for the “main” dialog, which was a scenario chosen randomly from level 1 through level 3. Finally, the subject was asked to answer a questionnaire.

The average time from registration to filling out the questionnaire was 15-20 minutes. Since we conducted 4-5 hours of experiments each day, excluding system setup, we were able to obtain dialog data for 15 subjects per day.

Table 10 is an overview of FED data.

Table 10. Overview of FED

	J (to E)	E (to J)	J (to C)	C (to J)
Number of utterances	608	660	344	484
Number of speakers	7	39	6	36
Number of word tokens	3,851	4,306	2,017	422
Number of word types	727	668	436	382

6.6 Performance Evaluation

We collected questionnaires from all subjects. As mentioned above, all of the Chinese-speaking subjects were college students. Therefore, they had at least a basic understanding of Japanese because they attend lectures given in Japanese. Therefore, in the following, we will focus on the English side.

First, overall performance is measured in terms of subjective scores from A to D, defined as follows.

(A) Perfect: no problems in either information or grammar.

(B) Good: easy to understand, with either some unimportant information missing or flawed grammar.

(C) Fair: broken but understandable with effort.

(D) Nonsense or no-output: (including ASR errors).

Table 11 shows results of English-Japanese translation. About 50% of the utterances were translated into the target language with their original meaning preserved (*e.g.*, at rank B or above).

Table 11. Results of English-Japanese translation

Rank	J to E (%)	E to J (%)
A	37.1	36.2
B	10.2	18.2
C	10.9	5.7
D	41.4	24.5

The ultimate goal of speech translation systems is to help users achieve their conversational goals. Instead of evaluating “goal achievement,” we asked them to subjectively evaluate during the course of conversations to what extent 1) they could understand their partner’s utterances, and 2) they felt that their utterances were correctly understood. Table 12 shows the questionnaire results on these issues.

Table 12. Results of questionnaires on understanding a partner's utterances (English side)

	Make the hearer understood (%)	Understood what the partner said (%)
Complete	8.3	22.2
Almost	41.6	50.0
Half	33.3	22.2
Little	16.7	5.6

Note that, although the number of subjects (*i.e.*, samples) is limited, the table does show that roughly half the subjects felt they could almost understand and make themselves understood by their partners. The result seems to coincide with the overall performance shown in Table 11.

7. Conclusion

This paper described our experience with multilingual spoken language corpus development at our research institution, focusing in particular on speech recognition and natural language processing for speech translation of travel conversations.

First, we introduced an interpreter-aided spoken dialog corpus called SLDB, and mentioned corpus configuration. Next, we introduced BTEC, which was built to train machine translation of spoken language among Japanese, English, and Chinese speakers. BTEC and SLDB are designed to be complementary. BTEC is a collection of Japanese sentences and their translations, and SLDB is a collection of transcriptions of bilingual spoken dialogs. Whereas BTEC covers a wide variety of travel domains, SLDB covers a limited domain, *i.e.*, hotel situations. BTEC contains approximately 588k utterance-style expressions, and SLDB contains about 16k utterances.

Finally, we discussed a multilingual spoken dialog corpus between Japanese, English, and Chinese created using speech-to-speech translation systems. MAD was developed as a development corpus and we presented both BTEC and SLDB can be used to handle with MAD-type tasks. FED was planned as the evaluation corpus. According to analysis of the questionnaire, roughly half the subjects felt they could understand and make themselves understood by their partners.

In the future, we plan to expand our activities to multilingual spoken language communication research and development involving both verbal and nonverbal communication. Information is available at the following URL: http://www.atr.jp.

Acknowledgments

The work reported here was mainly conducted at ATR Spoken Language Communication Research Laboratories. The authors are grateful to Prof. Seiichi Yamamoto, Dr. Satoshi Nakamura, and all other staff who helped construct corpora.

The FED corpus was collected within the framework of “Social experiments on supporting non-Japanese-speaking tourists using information technologies”, which was carried out by the Osaka prefectural government and related offices in the winter of 2004. In these experiments, Osaka’s prefectural government negotiated with management of facilities frequented by foreign tourists, such as airports and bus terminals, to provide the necessary assistance (*e.g.*, use of public space and electricity). The government also gathered the volunteer subjects.

References

Höge, H., “Project Proposal TC-STAR: Make Speech to Speech Translation Real,” *Proc. of International Conference on Language Resources and Evaluation*, 2002, pp. 136–141.

Itoh, G., Ashikari, Y., Jitsuhiro, T., and Nakamura, S., “Summary and Evaluation of Speech Recognition Integrated Environment ATRASR,” *Proc. of Autumn Meeting of the Acoustical Society of Japan*, 1-P-30, 2004, pp. 221–222.

IWSLT, *Proc. of International Workshop on Spoken Language Translation,* Kyoto, Japan, 2006.

Kawai, H., Toda, T., Ni, J., Tsuzaki, M., and Tokuda, K., “XIMERA: a New TTS from ATR Based on Corpus-based Technologies,” *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004, pp. 179–184.

Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., “Creating Corpora for Speech-to-Speech Translation,” *Proc. of European Conference on Speech Communication and Technology*, 2003, pp. 381–382.

Kikui, G., Yamamoto, S., Takezawa, T., and Sumita, E., “Comparative Study on Corpora for Speech Translation,” *IEEE Trans. on Audio, Speech, and Language Processing*, 14 (5), 2006, pp. 1674–1682.

Lavie, A., A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavaldà, T. Zeppenfeld, and P. Zhan, “JANUS-III: Speech-to-speech Translation in Multiple Language,” *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 99–102.

Lazzari, G., “TC-STAR: a Speech to Speech Translation Project,” *Proc. of International Workshop on Spoken Language Translation*, 2006, pp. xiv–xv.

Mizushima, M., T. Takezawa, and G. Kikui, “Effects of Audibility of Partner’s Voice and Visibility of Translated Text in Machine-Translation-Aided Bilingual Spoken

Dialogues," *IPSJ SIG Technical Reports*, 2004 (74), 2004-HI-109-19/2004-SLP-52-19, 2004, pp. 99–106.

Morimoto, T., T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu, "ATR's Speech Translation System: ASURA," *Proc. of European Conference on Speech Communication and Technology*, 1993, pp. 1291–1294.

Morimoto, T., N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki, "A Speech and Language Database for Speech Translation Research," *Proc. of International Conference on Spoken Language Processing*, 1994, pp. 1791–1794.

Rayner, M., I. Bretan, D. Carter, M. Collins, V. Digalakis, B. Gambäck, J. Kaja, J. Karlgren, B. Lyberg, S. Pulman, P. Price, and C. Samuelsson, "Spoken Language Translation with Mid-90's Technology: a Case Study," *Proc. of European Conference on Speech Communication and Technology*, 1993, pp. 1299–1302.

Roe, D. B., P. J. Moreno, R. W. Sproat, F. C. N. Pereira, M.D. Riley, and A. Macarrón, "A Spoken Language Translator for Restricted-domain Context-free Languages," *Speech Communication*, 11, 1992, pp. 311–319.

Roukos, S., "Recent Results on MT Evaluation in the GALE Program," *Proc. of International Workshop on Spoken Language Translation*, 2006, pp. xvi–xvii.

Sumita, E., H. Nakaiwa, and S. Yamamoto, "Corpus-Based Translation Technology for Multi-lingual Speech-to-Speech Translation," *Proc. of Spring Meeting of the Acoustical Society of Japan*, 1-8-26, 2004, pp. 57–58.

Takezawa, T., T. Morimoto, and Y. Sagisaka, "Speech and Language Databases for Speech Translation Research in ATR," *Proc. of Oriental COCOSDA Workshop*, 1998, pp. 148–155.

Takezawa, T., F. Sugaya, M. Naito, and S. Yamamoto, "A Comparative Study on Acoustic and Linguistic Characteristics Using Speech from Human-to-Human and Human-to-Machine Conversations," *Proc. of International Conference on Spoken Language Processing*, III, 2000, pp. 522–525.

Takezawa, T., E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World," *Proc. of International Conference on Language Resources and Evaluation*, 2002, pp. 147–152.

Takezawa, T. and G. Kikui, "Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation," *Proc. of European Conference on Speech Communication and Technology*, 2003, pp. 2757–2760.

Takezawa, T., A. Nishino, K. Takashima, T. Matsui, and G. Kikui,, "An Experimental System for Collecting Machine-Translation-Aided Dialogues," *Proc. of Forum on Information Technology*, E-036, 2003, pp. 161–162.

Takezawa, T. and G. Kikui, “A Comparative Study on Human Communication Behaviors and Linguistics Characteristics for Speech-to-Speech Translation,” *Proc. of International Conference on Language Resources and Evaluation*, 2004, pp. 1589–1592.

Wahlster, W. (Ed.), “Verbmobil: Foundations of Speech-to-Speech Translation,” Springer, Germany, 2000.

Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 325-348

Exploiting Pinyin Constraints in Pinyin-to-Character Conversion Task: a Class-Based Maximum Entropy Markov Model Approach

Jinghui Xiao[*], Bingquan Liu[*], and Xiaolong Wang[*]

Abstract

The Pinyin-to-Character Conversion task is the core process of the Chinese pinyin-based input method. Statistical language model techniques, especially ngram-based models, are mostly adopted to solve that task. However, the ngram model only focuses on the constraints between characters, ignoring the pinyin constraints in the input pinyin sequence. This paper improves the performance of the Pinyin-to-Character Conversion system through exploitation of the pinyin constraints. The MEMM framework is used to describe the pinyin constraints and the character constraints. A Class-based MEMM (C-MEMM) model is proposed to address the MEMM efficiency problem in the Pinyin-to-Character Conversion task. The C-MEMM probability functions are strictly deduced and well formulized according to the Bayes rule and the Markov property. Both the cases of hard class and soft class are well discussed. In the experiments, C-MEMM outperforms the traditional ngram model significantly by exploitation of the pinyin constraints in the Pinyin-to-Character Conversion task. In addition, C-MEMM can well utilize the syntax and semantic information in word class and further improve the system performance.

Keywords: Pinyin-to-Character Conversion, MEMM, Class-Based

1. Introduction

The standard keyboard was initially designed for native English speakers. In Asia, such as China, Japan and Thailand, people cannot input their language through the standard keyboard directly. Asian text input becomes one of the challenges for computer users in Asia. Therefore, an Asian language input method is one of the most difficult problems in Asian language processing.

[*] School of Computer Science and Techniques, Harbin Institute of Technology, Harbin, 150001, China
E-mail: {xiaojinghui, liubq, wangxl}@insun.hit.edu.cn

[Received October 12, 2006; Revised May 15, 2007; Accepted May 23, 2007]

For Chinese, the input methods can be roughly divided into two types: one is the structure-based or shape-based input method, which was developed based on the structure of Chinese characters, such as the Wubi method [Wang 2005], Cangjie method, Boshiamy method, among others. These methods can reach a high input speed by a skilled user. However, a lot of effort is required to master them. The other is the pronunciation-based input method, such as the Insun input method [Wang 1993], Microsoft input method, Bopomofo, among others. These methods are easy to learn. The user can input the Chinese character with scarcely any training, on the condition that they can pronounce it correctly. Hybrid input methods have also been proposed, *i.e.* Renzhi and Tze-loi input method. However, they only possess a limited share of the market.

The Pinyin-based input method is one of the most important pronunciation-based input methods. Pinyin is a system of Romanization for standard Mandarin. It uses Roman letters to represent the sound of Chinese characters. Hanyu Pinyin is the most common variant of pinyin in use. It was approved in 1958, and the government of the People's Republic of China has adopted Hanyu pinyin as the phonetic instruction in the mainland of China. In 1979, Hanyu pinyin was adopted by the International Organization for Standardization (ISO) as the standard Romanization for modern Chinese [ISO 7098: 1991]. The Pinyin-based input method dominates the market of Chinese input methods. It is said that over 97% of Chinese computer users are using pinyin to input Chinese [Chen 1997].

According to the scale of input unit, the pinyin-based input method can be divided into three types: the character-level input method, the word-level or phrase-level input method, and the sentence-level input method, respectively. The sentence-level input method usually achieves higher accuracy by exploitation of more context information than the other two. It has become the most prevalent pinyin-based input method. The Pinyin-to-Character Conversion task aims to convert a sequence of pinyin strings into one Chinese sentence. It is the core process of the sentence-level pinyin-based input method. Therefore, improving the performance of the Pinyin-to-Character Conversion system is well worth studying. In addition, the Pinyin-to-Character Conversion task can be taken as a simplified task of automatic speech recognition, both of which aim to convert phonetic information into character sequence. However, the Pinyin-to-Character Conversion task doesn't have to deal with acoustic ambiguity because the pinyin strings are directly input through the keyboard. Therefore, the technique is also illuminative in the task of automatic speech recognition.

The linguist approach [Wang 1993; Hsu and Chen 1993; Kuo 1995] and the statistical approach [Zhang *et al*. 1998; Xu *et al*. 2000; Wu 2000; Gao *et al*. 2002; Gao *et al*. 2005; Xiao *et al*. 2005a] are two technical approaches to the Pinyin-to-Character Conversion task. The statistical approach is mainly based on the technique of statistical language models, especially the ngram model and its variant forms. In recent years, it has drawn great interest due to its efficiency and robustness. However, several drawbacks have also been found in the traditional ngram model. First, according to Zipf's law [Zipf 1935], there are a lot of words which rarely or never occur in the training corpus. The data sparseness problem is severe [Brown *et al.* 1992] in the ngram model. Second, long distance constraints are difficult to capture since the ngram model only focuses on local lexical constraints. Third, it's hard to utilize the linguistic knowledge of the ngram model.

Many techniques have been proposed to address the drawbacks of the traditional ngram model. To solve the data sparseness problem, various kinds of smoothing techniques have been proposed, such as additive smoothing [Jeffreys 1948], Katz smoothing [Katz 1987], linear interpolation smoothing [Jelinek and Mercer 1980], semantic based smoothing [Xiao *et al.* 2005b; Xiao *et al.* 2006]. To utilize the linguistic knowledge, a set of linguistic rules are generated automatically and they are incorporated into the traditional ngram model by a hybrid ngram model [Wang *et al.* 2005]. Hsu [Hsu 1995] proposes the context sensitive model (CSM) in which the semantic patterns are captured by the templates. As much as 96% accuracy, which is the best result of the traditional Chinese input methods as far as we know, is reported for CSM on the Phoneme-to-Character Conversion task. Trigger techniques have been proposed [Zhou and Lua 1998] and word-pair techniques have been proposed [Tsai and Hsu 2002; Tsai *et al.* 2004; Tsai 2005; Tsai 2006]. The linguist knowledge can be effectively described by triggers and pairs; meanwhile, the long distance constraints can be well captured. Compared with the commercial input system (MS-IME 2003), effective improvements have been achieved by these techniques [Tsai 2006]. Wang [Wang *et al.* 2004] utilizes the theory of rough set so as to discover the linguistic knowledge and incorporate it into the Pinyin-to-Character Conversion system. Compared with the traditional ngram model, Wang 's system achieves a higher accuracy with a smaller storage requirement. Xiao [Xiao *et al.* 2005a] incorporates the word positional information into the Pinyin-to-Character Conversion system and achieves encouraging results in experiments. Gao [Gao *et al.* 2005] proposes the Minimum Sample Risk (MSR) principle to estimate the parameters of the ngram model. Success has been achieved with this principle for a Japanese input method.

What's more, some techniques have been proposed especially for Chinese text input method. A Pinyin-to-Character Conversion system with spelling-error correction was developed by Zhang [Zhang *et al.* 1997]. In the system, a rule-based model is designed to correct typing errors when the user inputs pinyin strings. Not only can the system accept the correct pinyin input, but it can also tolerate common typing errors. Similar work has been done by Chen [Chen and Lee 2000]. Chen constructs a statistical model to correct user typing errors. Moreover, Chen proposes a modeless input technique in which the user can input English using a Chinese input method, not requiring language mode switch.

However, there is another drawback of the ngram model in the Pinyin-to-Character

Conversion task, which has been ignored by most researchers. It takes no account of pinyin constraints on the input pinyin sequence while actually in the process of Pinyin-to-Character Conversion. This paper regards that the pinyin information from the pinyin sequence is helpful for selecting the correct character sequence in the Pinyin-to-Character Conversion task. First, the current input pinyin string is helpful for selecting the correct character which corresponds to that pinyin. For example, the input pinyin sequence is "ta1 shi4 di2 shi4 you3?" which should be converted into "他是敌是友?" ("Is he an enemy or friend?"). Let's focus on the third pinyin string of "di2". There are two homonyms which correspond to it: "敌" and "的". (There are actually many homonyms, but let 's only focus on "敌" and "的" for simplification). "的" is one of the most frequent Chinese characters and its frequency is usually much higher than "敌". According to the ngram model, the above pinyin sequence should be converted into "他是的是友？" which is a wrong conversion. However, "的" is a polyphone which corresponds to both "di2" and "de5". In Chinese, "的" is usually pronounced as "de5" instead of "di2". ("的" is pronounced as "di2" only in the word "的确" (certainly)). The frequency of "的" mainly comes with its pronunciation "de5". If the pinyin information is considered in the above conversion, the co-occurrences of "的" and "di2" are usually lower than that of "敌" and "di2". Then, the above pinyin sequence is correctly converted into "他是敌是友?". Second, the contextual information, especially the future information, can be well exploited in the pinyin constraints. For example, there are two pinyin sequences. The first one is "yi4 zhi1 ke3 ai4 de5 xiao3 hua1" which should be converted into "一枝可爱的小花" (This is a lovely flower). The second pinyin "zhi1" should be converted into "枝" which is determined by its future character "花" (flower). The second pinyin sequence is "yi4 zhi1 ke3 ai4 de5 xiao3 hua1 mao1" which should be converted into "一只可爱的小花猫" (This is a lovely cat). The second pinyin "zhi1" should be converted into "只" which is determined by its future character "猫" (cat). However, according to the ngram model, the conversion of "zhi1" is only determined by its history information which is the character "一" in the above two cases. The characters of "花" and "猫" are both the further information that the ngram model can not exploit. Therefore, the same probabilities are assigned to both the characters of "只" and "枝". They can not be distinguished by the ngram model. In the above two conversions, at least one of them would be converted incorrectly. However, if the pinyin constraints are considered, the constraints of "hua1" and "mao1", which correspond to the characters of "花" and "猫", are exploited and imposed on the conversion of "zhi1". Then, the above two cases can be distinguished and the correct conversions can be obtained. Third, the long distance constraints can be exploited from the pinyin sequence. As for the ngram model, it has to construct a high-order model to capture the long distance constraints. However, high-order ngram models suffer from the curse of dimensionality which usually leads to a severe data sparseness problem. The current model order is usually 2 or 3. In the above example, in order to exploit the constraints of “花” and “猫” on the conversion of “zhi1”, it has to build up at least a 7-order ngram model which suffers from a great data sparseness problem and cannot work well in reality. However, the pinyin constraints are collected as features and exploited under the Maximum Entropy (ME) framework in this paper. The context window size can be relatively large (*i.e.* 5 pinyin strings or 7 pinyin strings) without the curse of dimensionality. Then the constraints of “花” and “猫” can be imposed on the conversion of “zhi1” by exploitation of their pinyin information.

This paper aims to improve the performance of the Pinyin-to-Character Conversion system by exploitation of the pinyin constraints from pinyin sequence. The pinyin constraints are described under the ME framework [Berge *et al.* 1996], and the character constraints are modeled by the traditional ngram model. Combining these two models into a unified framework, the paper builds the Pinyin-to-Character Conversion system on a MEMM model [McCallum *et al.* 2000]. However, the label set on the Pinyin-to-Character Conversion task is the Chinese lexicon. The scale of Chinese lexicon is usually in the range of $10^4 \sim 10^6$, which is too large for the current training algorithms of MEMM. Therefore MEMM cannot be directly applied to the Pinyin-to-Character Conversion task. This paper involves the addition of the class of target label into traditional MEMM and proposes a Class-based Maximum Entropy Markov Model (C-MEMM) so as to solve the MEMM efficiency problem in the Pinyin-to-Character Conversion task. In C-MEMM, the pinyin constraints are first imposed on the class sequence instead of the target label sequence as MEMM does. The classes of target label can be obtained by some automatic algorithms [Li 1998; Chen and Huang 1999; Gao *et al.* 2001] or from some pre-defined thesauri [Mei *et al.* 1983]. The scale of class set is usually much smaller than that of target label, which makes it feasible to train C-MEMM under the Maximum Entropy principle. Then, these constraints are conveyed from the class sequence to the target label sequence. So, C-MEMM can efficiently exploit the pinyin constraints from pinyin sequence and get effective improvement in the Pinyin-to-Character Conversion task.

The paper is organized as follows: the MEMM model is briefly reviewed in Section 2. In Section 3, the C-MEMM model is proposed and its probability functions are deduced according to the Bayes rule and the Markov property. Both the cases of hard class and soft class are discussed in detail. Experimental results and discussions are provided in Section 4. The related works are described in Section 5, and the conclusions are drawn in Section 6.

2. Brief Review of MEMM

MEMM is a powerful tool used to perform the sequence labeling task, which is to determine a state sequence according to the observation sequence. Different from the ngram model, MEMM not only makes use of the constraints between states but also utilizes the constraints from observations. MEMM integrates these two kinds of constraints into a uniform conditional probability function. More formally, given the observation sequence of $O = o_1, o_2 ... o_n$ and the state sequence of $S = s_1, s_2 ... s_n$, MEMM estimates the conditional probability of $P(S \mid O)$. The probability function of MEMM can be deduced in the following way:

$$\begin{aligned}
P(S \mid O) &= p(s_1, s_2 ... s_n \mid o_1, o_2 ... o_n) \\
&\overset{Bayesian\ Rule}{=} p(s_1 \mid o_1, o_2 ... o_n) p(s_2, s_3 ... s_n \mid o_1, o_2 ... o_n, s_1) \\
&\overset{Markov\ Property}{=} p(s_1 \mid o_1) p(s_2, s_3 ... s_n \mid o_1, o_2 ... o_n, s_1) \\
&\overset{Bayesian\ Rule}{=} p(s_1 \mid o_1) p(s_2 \mid o_1, o_2 ... o_n, s_1) p(s_3 ... s_n \mid o_1, o_2 ... o_n, s_1, s_2) \\
&\overset{Markov\ Property}{=} p(s_1 \mid o_1) p(s_2 \mid s_1, o_2) p(s_3 ... s_n \mid o_1, o_2 ... o_n, s_1, s_2) \\
&\qquad \cdots\cdots\cdots \\
&= p(s_1 \mid o_1) \prod_{i=2}^{n} p(s_i \mid s_{i-1}, o_i)
\end{aligned} \tag{1}$$

MEMM estimates the probability of $p(s_i \mid s_{i-1}, o)$ under the ME principle so as to utilize the overlapping features. The ME principle assumes that the trained model should be consistent with certain constraints derived from the training data; meanwhile the model should make the fewest assumptions about the data. To predicate the current state s, the contextual information of s is extracted from the training data and represented as the feature function:

$$f(h, s) = \begin{cases} 1 & if\ h = h^* \ and\ s = s^* \\ 0 & otherwise \end{cases} \tag{2}$$

where h is the contextual information of state s and h^* (or s^*) is the concrete instance of h (or s). The following constraints are imposed so that the expectation of each feature in the learned model should be consistent with its empirical value in the training corpus. More formally, the constraints can be expressed as:

$$E_p(f) = E_{\tilde{p}}(f) \tag{3}$$

where $E_p(f)$ is the model expectation and is defined as:

$$E_p(f) = \sum_{h,s} \tilde{p}(h) p(s \mid h) f(h, s) \tag{4}$$

and $E_{\tilde{p}}(f)$ is the empirical expectation and is defined as:

$$E_{\tilde{p}}(f) = \sum_{h,s} \tilde{p}(h, s) f(h, s) \tag{5}$$

Under these constraints, ME principle guarantees a learned model as uniform as possible, and the model can be obtained by maximizing the conditional entropy of the training data:

$$H(p) = -\sum_{h,s} \tilde{p}(h)p(s \mid h)\log p(s \mid h) \tag{6}$$

It results in the probability function of exponential form:

$$p(s \mid s^{'}, o) = \frac{1}{Z(h, s^{'})}\exp(\sum_{i} \lambda_i f_i(h,s)) \tag{7}$$

where λ is the weight of the feature f_i, and Z is the normalization factor.

In the above formula, $p(s \mid s^{'}, o)$ associates observation with state transition, which makes data sparseness a serious problem. Therefore, a variant form of MEMM is proposed. It makes observations associated with state instead of state transition. Then, MEMM is decomposed into two sub-models: the ngram model and the ME model. The probability function is reformulated as:

$$p(s \mid s^{'}, o) = p(s \mid s^{'})p_{ME}(s \mid h) \tag{8}$$

where $p_{ME}(s \mid h)$ is the conditional probability which is estimated under the ME principle and has the exponential form. Since the data sparseness problem is prone to occur in the Pinyin-to-Character Conversion task, our work is based on Formula (8).

Accordingly, the training process of MEMM can be decomposed into two separate processes for the ngram model and the ME model. The ngram model can be effectively trained by the Maximum Likelihood Estimation (MLE) principle [Myung 2003]. For the ME model, there is no easy solution to get the optimal value of λ directly. Some iterative algorithms, *i.e.* the Generalized Iterative Scaling (GIS) algorithm [Darroch and Ratcliff 1972] and the Improved Iterative Scaling (IIS) algorithm [Pietra *et al.* 1997], are usually adopted. However, the time complexity of the iterative algorithm is far beyond the complexity of the MLE method, and it becomes the bottleneck of the training process of MEMM. When the scale is large, it is infeasible to use the iterative algorithm to train the MEMM model because of the high complexity.

3. Principle of Class-Based MEMM

This paper involves the class of state in traditional MEMM so as to address its efficiency problem on a large scale of state set. A Class-based MEMM model is proposed and its probability functions are strictly deduced and well formulized both in the case of hard class and soft class. The section is structured as follows. First, it presents C-MEMM in the case of hard class. Second, it describes C-MEMM in the case of soft class. Third, it provides ways to get the class of the state.

3.1 C-MEMM on Hard Class

The simplest way to construct C-MEMM is to substitute state with class of state in the probability of $p_{ME}(s \mid h)$ in Formula (8). Then, $p_{ME}(c \mid h)$ is used to simulate $p_{ME}(s \mid h)$ in which c is the class of s. However, the predicative capability of $p_{ME}(c \mid h)$ is much lower than that of $p_{ME}(s \mid h)$, which decreases the overall performance of C-MEMM. This paper begins the work from calculating the conditional probability of sequential data, and re-deduces the probability functions for C-MEMM according to the Bayes rule and the Markov property. More formally, the following notations are defined:

- $O = o_1, o_2 ... o_n$: the observation sequence
- $S = s_1, s_2 ... s_n$: the state sequence
- $C = c_1 c_2 ... c_n$: the class sequence which corresponds to *S*. It is unique in the case of hard class.

In the case of hard class, where the class sequence is completely determined by the state sequence, the following equation can be made:

$$P(S \mid O) = P(S, C \mid O) \; . \tag{9}$$

Then, the probability function of C-MEMM can be deduced through the following process:

$$P(S \mid O) = P(S, C \mid O) \overset{Bayesian\,Rule}{=} P(C \mid O) \times P(S \mid C, O) \; . \tag{10}$$

According to the Bayes rule, the conditional probability of sequential data is decomposed into two conditional probabilities. The probability of $P(C \mid O)$ can be further decomposed by the Bayes rule and the Markov property, exactly as the process of Formula (1). The ultimate formula is directly presented as below:

$$P(C \mid O) = p(c_1 \mid o_1) \prod_{i=2}^{n} p(c_i \mid c_{i-1}, o_i) = p(c_1 \mid o_1) \prod_{i=2}^{n} p(c_i \mid c_{i-1}) p_{ME}(c_i \mid h_i) \tag{11}$$

In the above formula, $p(c_i \mid c_{i-1}, o_i)$ is further decomposed by Formula (8).

For the probability of $P(S \mid C, O)$, the decomposing process is a little more complex and an additional independent assumption should be made.

$$
\begin{aligned}
P(S \mid C,O) &= p(s_1 s_2 \ldots s_n \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n) \\
&\overset{\text{Bayesian Rule}}{=} p(s_1 \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n) \times p(s_2 \ldots s_n \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n, s_1) \\
&\overset{\text{Markov Property}}{=} p(s_1 \mid c_1, o_1) \times p(s_2 \ldots s_n \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n, s_1) \\
&\overset{\text{Bayesian Rule}}{=} p(s_1 \mid c_1, o_1) \times p(s_2 \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n, s_1) \times p(s_3 \ldots s_n \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n, s_1 s_2) \\
&\overset{\text{Markov Property}}{=} p(s_1 \mid c_1, o_1) \times p(s_2 \mid c_2, o_2, s_1) \times p(s_3 \ldots s_n \mid c_1 c_2 \ldots c_n, o_1 o_2 \ldots o_n, s_1 s_2) \\
&\qquad \cdots\cdots\cdots \\
&= p(s_1 \mid c_1, o_1) \times \prod_{i=2}^{n} p(s_i \mid c_i, o_i, s_{i-1}) \\
&\overset{\text{Independent Rule}}{=} p(s_1 \mid c_1, o_1) \times \prod_{i=2}^{n} p(s_i \mid c_i, o_i) \times p(s_i \mid s_{i-1})
\end{aligned}
\tag{12}
$$

The fore part of the above deduction is exactly the same as the process in Formula (1). In the following part, such an assumption is made that the state transition probability is independent of the emission probability. The local conditional probability of $p(s_i \mid c_i, o_i, s_{i-1})$ is then decomposed into two probabilities: $p(s_i \mid s_{i-1})$ and $p(s_i \mid c_i, o_i)$, in which $p(s_i \mid s_{i-1})$ is the state transition probability and $p(s_i \mid c_i, o_i)$ is the class-based emission probability. To gain more insight, $p(s_i \mid c_i, o_i)$ can be rewritten as $p(s_i \mid o_i, c_i)$ which is the emission probability that is conditioned on the class of c_i. Together with the decompositions of Formulas (10) and (11), the ultimate form of the probability function of C-MEMM can be obtained, presented as below:

$$P(S \mid O) = p(c_1 \mid o_1) p(s_1 \mid c_1, o_1) \prod_{i=2}^{n} p(c_i \mid c_{i-1}) p_{ME}(c_i \mid h_i) p(s_i \mid c_i, o_i) p(s_i \mid s_{i-1}) \tag{13}$$

Until now, the conditional probability of sequential data has been decomposed into four kinds of local conditional probabilities. The probability of $p_{ME}(c_i \mid h_i)$ is estimated under the ME principle and it has the exponential form:

$$p_{ME}(c \mid h) = \frac{1}{Z(h)} \exp\left(\sum_i \lambda_i f_i(h,c)\right). \tag{14}$$

As the scale of c is much smaller than that of s, it needs shorter time for C-MEMM to estimate $p_{ME}(c_i \mid h_i)$ than $p_{ME}(s_i \mid h_i)$, which makes it feasible to apply C-MEMM in the tasks with large scale of state set, *i.e.* the Pinyin-to-Character Conversion task. The other three probabilities are estimated by the Maximum Likelihood Estimation (MLE) principle, presented as below:

$$p(c_i \mid c_{i-1}) = \frac{C(c_{i-1}, c_i)}{C(c_{i-1})} \tag{15}$$

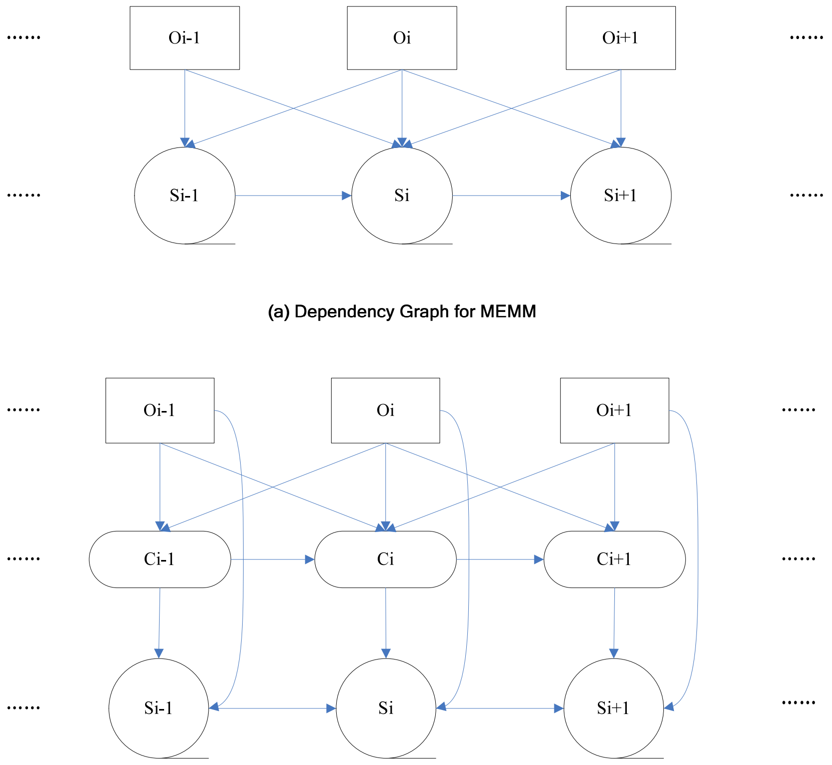
where $C(x)$ is the occurrence times of x in the training corpus.

$$p(s_i \mid s_{i-1}) = \frac{C(s_{i-1}, s_i)}{C(s_{i-1})} \tag{16}$$

$$p(s_i \mid c_i, o_i) = \frac{C(c_i, o_i, s_i)}{C(c_i, o_i)} \tag{17}$$

When applying C-MEMM in the Pinyin-to-Character Conversion tasks, the four kinds of local conditional probabilities are first estimated from the training corpus. Then, according to the input pinyin sequence, the probability of a character sequence candidate is calculated by Formula (13). Finally, the most probable character sequence is selected as the conversion results for the input pinyin sequence. Some dynamic programming algorithms can be utilized in the above process, *i.e*. the Viterbi algorithm.

In the remaining part of this section, the probability dependency graph in C-MEMM is presented and an intuitional description is provided on the functions of the four local conditional probabilities.



(a) Dependency Graph for MEMM

(b) Dependency Graph for C-MEMM

Figure 1. Probability Dependency Graphs for MEMM and C-MEMM

Presented as the above graphs, the constraints from the observation sequence are imposed directly on the state sequence in MEMM. The scale of the state set becomes a bottleneck in the training process of MEMM. However, in C-MEMM, there is a class sequence between the state sequence and the observation sequence. All the constraints from the observation sequence are imposed on the class sequence in C-MEMM, rather than directly on the state sequence as in MEMM. Since the scale of the class set is much smaller than that of the state set, the conditional probability of $p_{ME}(c_i \mid h_i)$, which connects the observation sequence with the class sequence, can be efficiently estimated under the ME principle. The constraints from the observation sequence are also well exploited by the probabilities of the class sequence. Furthermore, all the constraints from the observation sequence are conveyed from the class sequence into the state sequence by the conditional probabilities between these two sequences.

Concretely speaking, in Formula (13), the conditional probability of $p_{ME}(c_i \mid h_i)$ and the class transition probability of $p(c_i \mid c_{i-1})$ aim to model the constraints from the observation sequence and conserve them in the probability of the class sequence. The conditional probability of $p(s_i \mid c_i, o_i)$ conveys these constraints from the class sequence to the state sequence. The three conditional probabilities, together with the state transition probability of $p(s_i \mid s_{i-1})$, form the probability function of C-MEMM.

Moreover, since there is rich syntactic and semantic information in word class [Brown *et al.* 1992], C-MEMM used in the Pinyin-to-Character Conversion task can well utilize this additional information to realize further improvement.

3.2 C-MEMM on Soft Class

In the above section, the probability function of C-MEMM is deduced in the case of a hard class in which the state is restricted to only one class. However, in natural language processing tasks, *i.e.* the Pinyin-to-Character Conversion task, the state of C-MEMM is usually defined as word in the lexicon which usually belongs to multiple classes in nature. For example, part-of-speech (POS) can be taken as a natural hierarchy of word class. Most words possess more than one kind of POS tag. Each POS tag represents a certain syntactic and semantic property of the word. It is beneficial for C-MEMM to exploit all the properties of the word in natural language processing. The section studies C-MEMM in the case of soft class in which the state belongs to multiple classes. The probability function is re-deduced for C-MEMM.

In the case of a soft class, there are many class sequences corresponding to one state sequence. In order to calculate the probability of the state sequence, the conditional probabilities of all the class sequences should be summarized. Therefore, it is more complex to deduce the probability function of C-MEMM in the case of soft class than hard class. Similar to the case of the hard class, this section begins the work from calculating the conditional probability of sequential data, presented as below:

$$P(S \mid O) = \sum_{C} P(S, C \mid O) \overset{Bayesian\,Rule}{=} \sum_{C} P(C \mid O) \times P(S \mid C, O)\,. \tag{18}$$

The decompositions of $P(C \mid O)$ and $P(S \mid C, O)$ are exactly the same as those in the case of the hard class, which were presented in the above section. Then, the probability function in the case of soft class can be directly described as below:

$$P(S \mid O) = \sum_{c_1 \ldots c_n} \{p(s_1 \mid c_1, o_1) p(c_1 \mid o_1) \prod_{i=2}^{n} p(c_i \mid c_{i-1}) p_{ME}(c_i \mid h_i) p(s_i \mid c_i, o_i) p(s_i \mid s_{i-1})\}\,. \tag{19}$$

$p_{ME}(c_i \mid h_i)$, $p(c_i \mid c_{i-1})$, $p(s_i \mid c_i, o_i)$ and $p(s_i \mid s_{i-1})$ are estimated exactly in the same way as in the hard class. The probability dependency graph in the case of soft class is presented as below:

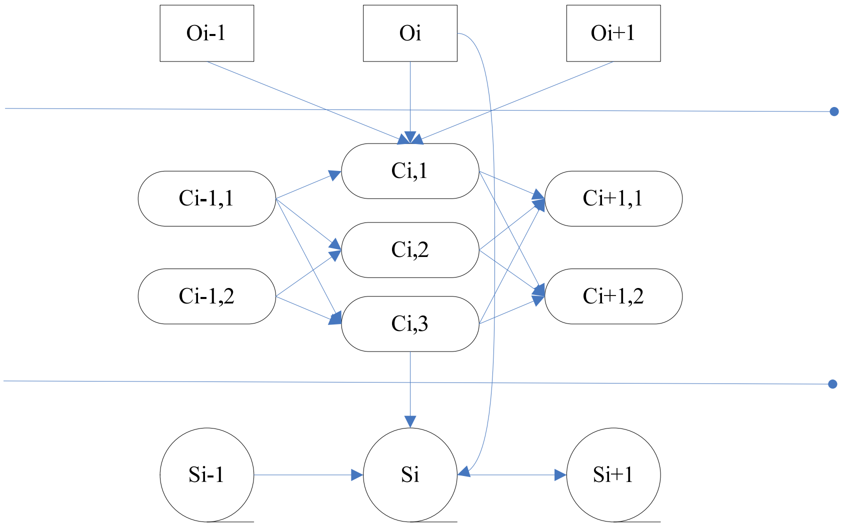


Figure 2. Probability Dependency Graph for C-MEMM on Soft Class

Differing from the case of a hard class, there are multiple class sequences between the observation sequence and the state sequence in the case of a soft class. In order to calculate the probability of $P(S \mid O)$, it is necessary to summarize all the conditional probabilities in these class sequences. The time complexity increases at an exponential rate with the length of sequence. Some dynamic algorithms, *i.e.* the forward algorithm and the backward algorithm, can calculate $P(S \mid O)$ efficiently at the polynomial time complexity. However, in the Pinyin-to-Character Conversion task, it is to find the optimal state sequence of S which maximizes the probability of $P(S \mid O)$. This is the *decoding problem of C-MEMM*. In a straightforward way, it's necessary to enumerate all the possible sequences of S and calculate the value of $P(S \mid O)$ for each sequence. The optimal sequence with the highest $P(S \mid O)$ is then selected from them. In reality, it is infeasible because of the high time complexity. The dynamic algorithm, *i.e*. the Viterbi algorithm, is expected to solve the decoding problem. However, Formula (19) makes a global summarization in the class sequences in which the Viterbi algorithm can not be applied. A simplification is then made in this paper. The global summarization, which is based on the whole sequence of class, is decomposed into the local summarization which is only based on the class at certain position. The probability function is simplified as below:

$$P(S \mid O) \approx \sum_{c_1} p(s_1 \mid c_1, o_1) p(c_1 \mid o_1) \times \prod_{i=2}^{n} \sum_{c_i} p(c_i \mid c_{i-1}) p_{ME}(c_i \mid h_i) p(s_i \mid c_i, o_i) p(s_i \mid s_{i-1}) \text{ .(20)}$$

The Viterbi algorithm can be applied to Formula (20) and can find the optimal state sequence of *S* in a polynomial time complexity. The dependency relationship graph is then described as below:

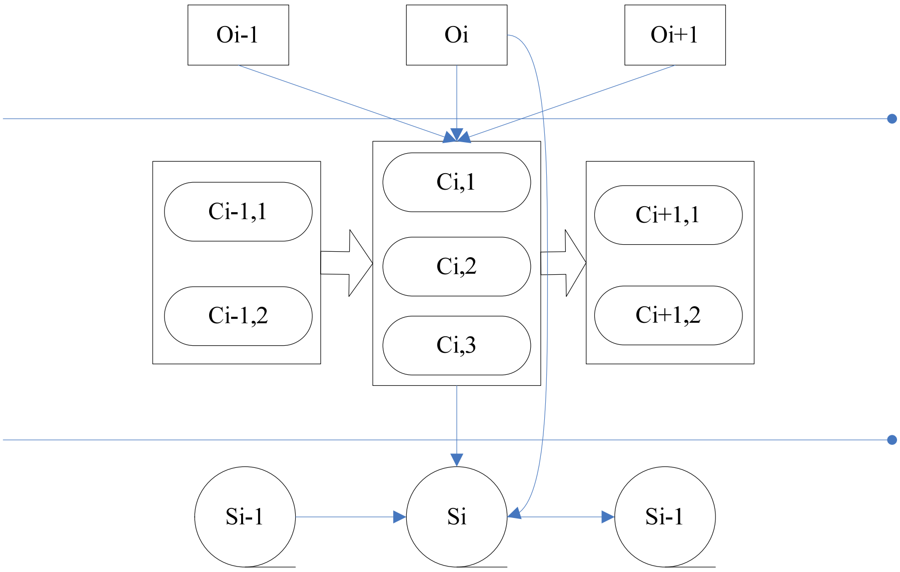


Figure 3. Probability Dependency Graph for the Simplified C-MEMM on Soft Class

3.3 Hierarchy of State Class

There are two ways to get the class of state. One is the statistical method, by which the state class is obtained by the clustering algorithm from the training corpus. However, according to Zip’s law, there are always low-frequency or zero-frequency states in the training corpus. Their frequencies are not statistically significant, and they can not be properly clustered by the statistical methods. The other method is getting the class from the pre-defined thesaurus. The hierarchy of class is defined by linguists according to the syntax and semantic information of each state. It can be taken as the well-defined hierarchy of state class. This paper attains the hierarchy of state class in the second way. TongyiciCilin is adopted as the hierarchy of state class in the case of hard class and the set of POS tag is adopted in the case of soft class. The detailed information is presented in Section 4.1.

4. EXPERIMENTS AND DISCUSSIONS

This section evaluates C-MEMM in the Pinyin-to-Character Conversion task. First, the data set is described. Second, the experimental results are presented. The performances of C-MEMM are evaluated both in the case of hard class and soft class. Third, the conclusion is drawn.

4.1 Data Set Description

This section describes the data set used in the experiments. First, information about the text corpus is presented. Then, the way to get pinyin corpus is described. Finally, the hierarchies of word class are presented.

Text Corpus

This paper chooses six months of the People's Daily corpus in 1998 as the text corpus in the experiments. The corpus has been annotated by Peking University with the POS tags and the name entities [Yu *et al*. 2003]. It has become the standard corpus in Chinese language processing in recent years [Emerson 2005]. There are 46 kinds of POS tag in the POS set. They are listed in Table 1:

Table 1. POS Set of Peking University

POS Set of Peking University							
Ag	*a*	*ad*	*an*	*Bg*	*b*	*c*	*Dg*
d	*e*	*f*	*g*	*h*	*i*	*j*	*k*
l	*Mg*	*m*	*Ng*	*n*	*nr*	*ns*	*nt*
nx	*nz*	*o*	*p*	*Qg*	*q*	*Rg*	*r*
s	*Tg*	*t*	*Ug*	*u*	*Vg*	*v*	*vd*
vn	*w*	*x*	*Yg*	*y*	*z*		

The text corpus is divided into two parts: the training corpus which consists of the first five months' corpora, and the testing corpus which is the sixth month's corpus. The detailed information is presented in Table 2:

Table 2. Description of the Text Corpus

	Training Corpus	Testing Corpus
Number of months	*1-5 months*	*6^{th} month*
Number of characters	*9.09×10^6*	*1.88×10^6*

Pinyin Corpus

The pinyin corpus is necessary for evaluating C-MEMM in the Pinyin-to-Character Conversion task. When C-MEMM is evaluated, the pinyin corpus is first converted into the character corpus by C-MEMM. Then, the conversion results are compared with the standard text corpus and the error rate is calculated. The pinyin corpus is obtained from the text corpus by a conversion toolkit[1] which achieves 99.7% accuracy on a golden pinyin corpus. In the experiments, the errors in the pinyin corpus could lead to the conversion error of C-MEMM. Therefore, the actual error rate of C-MEMM is a little lower than the reported results in this paper. However, there are not many errors in the pinyin corpus because of the high precision of our conversion toolkit. Thereby, the experimental results can be regarded to be close enough to the actual performance of C-MEMM.

Hierarchy of Word Class

Moreover, word class is necessary for building up C-MEMM in the Pinyin-to-Character Conversion task. The paper gets the hierarchy of word class from the compiled thesaurus which contains the word class information.

TongyiciCilin [Mei *et al.* 1983] is adopted as the hierarchy of word class in the experiments of hard class. TongyiciCilin was initially complied in 1982. There were initially 5.38×10^4 words which were organized into a tree structure according to their syntax and semantic information. The structure is shown in Figure 4. There are a total of four layers in the tree. The word is represented by the leaf node in the leaf layer. The word class is represented by the internal node in the internal layer. There is a road from each leaf node to the root node. On the road, there are several internal nodes of different layers which represent the classes of different scales that the leaf node belongs to. The node in the higher layer represents the class of bigger scale which usually corresponds to a more general concept of Chinese, and vice-versa. Each layer represents a pattern of word class, and the nodes in the same layer describe a way to cluster the words in TongyiciCilin. Moreover, the lower the layer is, the finer the word classes are, therefore the more syntactic and semantic information the layer contains. For example, the 3rd layer contains more syntactic and semantic information than the 1st layer does in Figure 4.

[1] http://www.insun.hit.edu.cn/product/viewproduct.asp?id=105

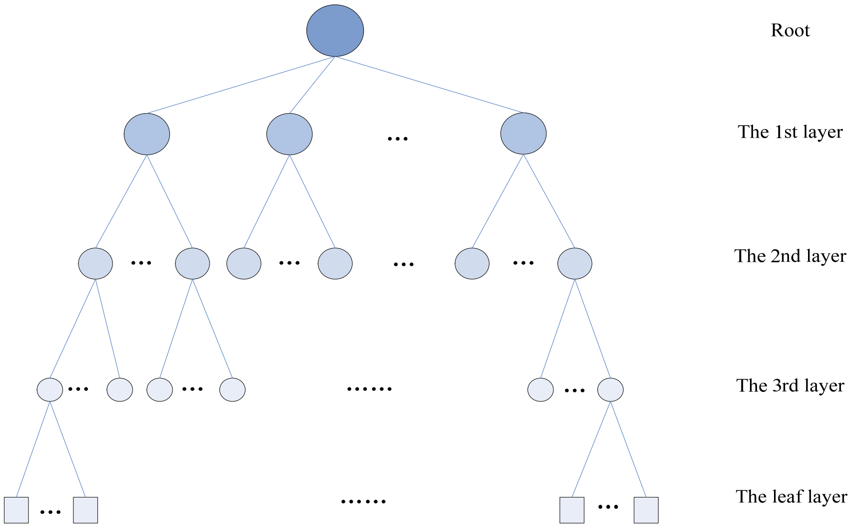


Figure 4. Hierarchy of Word Class in TongyiciCilin

In recent years, an extended version [Liu *et al*. 2005] of the original TongyiciCilin has been complied. Some infrequent words have been deleted, while some new words have been added. The scale of the lexicon in the new version is up to 7.73×10^4. The detailed information is described in Table 3:

Table 3. Description of TongyiciCilin (new version)

Description of TongyiciCilin	
Scale of lexicon	*7.73×10^4*
Number of Cluster in 1^{st} layer	*12*
Number of Cluster in 2^{nd} layer	*97*
Number of Cluster in 3^{rd} layer	*1428*

This paper adopts the new version of TongyiciCilin in the experiments of hard class.

In the experiments of soft cluster, the POS set is a natural choice for the hierarchy of word class. The information of the POS set has been provided in the beginning of this section.

4.2 Experiments on the Hard Class

This section investigates C-MEMM in the case of hard class in the Pinyin-to-Character Conversion task. TongyiciCilin is adopted as the hierarchy of word class. All the words in TongyiciCilin are adopted as the lexicon. The bigram model is taken as the baseline model. The additive smoothing technique is utilized. One order of C-MEMM is evaluated. Ten feature types of the pinyin constraints are extracted and exploited in C-MEMM. They are listed in Table 4:

Table 4. Feature Types in C-MEMM

	Feature Type	Feature Description
Atomic Feature Type	Yin_i	*The current pinyin*
	Yin_{i-1}	*The previous pinyin*
	Yin_{i-2}	*The previous but one pinyin*
	Yin_{i+1}	*The next pinyin*
	Yin_{i+2}	*The next but one pinyin*
	$YinComb_i$	*The pinyin combination of the current word which usually consists of several pinyin strings.*
Combined Feature Type	$Yin_i\ Yin_{i-1}$	*The combination of* Yin_i *and* Yin_{i-1}
	$Yin_i\ Yin_{i+1}$	*The combination of* Yin_i *and* Yin_{i+1}
	$Yin_{i-1}\ Yin_{i-2}$	*The combination of* Yin_{i-1} *and* Yin_{i-2}
	$Yin_{i+1}\ Yin_{i+2}$	*The combination of* Yin_{i+1} *and* Yin_{i+2}

From the above feature types, two feature templates are constructed so as to investigate the effectiveness of different feature types in C-MEMM performances. In template one, the size of the context window is set to 3, based on which the model of C-MEMM-1 is constructed. In template two, the size of the context window is set to 5, based on which the model of C-MEMM-2 is constructed. The information is presented in Table 5:

Table 5. Feature Templates in C-MEMM

Feature Template	Feature Types	Model
Template One	Yin_i, Yin_{i-1}, Yin_{i+1}, $YinComb_i$, $Yin_i\ Yin_{i-1}$, $Yin_i\ Yin_{i+1}$	*C-MEMM-1*
Template Two	Yin_i, Yin_{i-1}, Yin_{i-2},Yin_{i+1}, Yin_{i+2}, $YinComb_i$, $Yin_i\ Yin_{i-1}$, $Yin_i\ Yin_{i+1}$, $Yin_{i-1}\ Yin_{i-2}$, $Yin_{i+1}\ Yin_{i+2}$	*C-MEMM-2*

As mentioned above, there are several ways to cluster a word in TongyiciCilin, each corresponding to an internal layer in the tree structure of TongyiciCilin. C-MEMM is built up based on each pattern of word class used separately for each internal layer in TongyiciCilin. The performance of C-MEMM is investigated and the error rates are presented in Table 6:

Table 6. Error Rate of C-MEMM in the case of Hard Class

	No cluster	Clusters of 1st layer	Clusters of 2nd layer	Clusters of 3rd layer
Baseline	9.15%	---	---	---
C-MEMM-1	---	6.10%	**5.84%**	5.85%
Reduction	---	33.33%	**36.17%**	36.07%
C-MEMM-2	---	5.73%	5.46%	**5.28%**
Reduction	---	37.38%	40.33%	**42.30%**

The error rate of the baseline model is presented in the category of ‘No cluster’ from which the error rate reductions are calculated. According to the experimental results, C-MEMM outperforms the baseline model significantly with great error rate reduction. As much as 36.17% reduction has been achieved by C-MEMM-1 and 42.30% reduction has been yielded by C-MEMM-2. It proves that the predicative capability of C-MEMM is superior to that of the ngram model in the Pinyin-to-Character Conversion task. In addition, comparing the performance of C-MEMM-1 with C-MEMM-2, C-MEMM-2 outperforms C-MEMM-1 slightly, due to modeling the richer feature types of the pinyin constraints. This fact proves that the improvements of C-MEMM in the Pinyin-to-Character Conversion task are due to the exploitation of the pinyin constraints from the input pinyin sequence. Finally, the section investigates the performance of C-MEMM based on different patterns of word class. As mentioned in the above section, there is an increase of syntactic and semantic information contained in the word classes from the 1st internal layer to the 3rd internal layer of TongyiciCilin. From Table 6, it can be found that the error rates of C-MEMM generally decrease from the 1st layer to the 3rd layer, which proves that C-MEMM can make good use of the syntactic and semantic information from the word classes and attain further improvement.

To draw a conclusion, C-MEMM achieves significant error rate reductions from the ngram model in the Pinyin-to-Character Conversion task by exploitation of pinyin constraints. In addition, C-MEMM makes good use of the syntactic and semantic information in word class and sees further improvement.

4.3 Experiments on the Soft Class

This section evaluates C-MEMM in the case of soft class. The POS set of Peking University is taken as the hierarchy of word class. A word list compatible with the POS set is adopted as the lexicon. Other settings are the same as those in the case of hard class. The experimental results are presented in Table 7:

Table 7. Error Rate of C-MEMM in the case of Soft Class

	Baseline	C-MEMM-1	C-MEMM-2
Error rate (%)	8.37%	6.00%	5.82%
Reduction (%)	------	28.15%	30.47%

The experimental results are similar to the results in the case of hard class. First, C-MEMM outperforms the baseline model significantly. As much as 28.15% error rate reduction was achieved by C-MEMM-1 and a 30.47% error rate reduction was obtained by C-MEMM-2. This proves that C-MEMM is much more powerful than the ngram model. Second, C-MEMM-2 gets better performance than C-MEMM-1, due to modeling the richer feature types of the pinyin constraints. This indicates that the improvements of C-MEMM are due to the exploitation of the input pinyin information. Therefore, the conclusion is drawn that C-MEMM (hard-class based or soft-class based) improves the performance of the Pinyin-to-Character Conversion system significantly by exploitation of the pinyin constraints from the pinyin sequence.

In the remaining part of this section, the performance of the soft-class based MEMM is compared with the hard-class based MEMM. However, the experimental results in this section can not be compared directly with the results in Section 4.2, due to the fact that different lexica were used in the two sections. For fair comparison, a hierarchy of hard class is created from the hierarchy of soft class in this section. It restricts only one POS tag for each word in the lexicon. The most frequent POS tag of that word is adopted in the hierarchy of hard class. The experimental results are presented in Table 8:

Table 8. Comparison between Soft-class based MEMM and Hard-class based MEMM

	Baseline	C-MEMM-1	C-MEMM-2
No class	8.37%	------	------
Hard class	------	6.21%	6.17%
Soft class	------	6.00%	5.82%

As shown in Table 8, the soft-class based MEMM performs better than the hard-class based MEMM to some extent, proving that the soft-class based MEMM can exploit the comprehensive properties of word to achieve better performance.

4.4 Comparison with Class-based Ngram Model

The class-based ngram model enhances the traditional ngram model by involving word class [Brown *et al*. 1992]. The data sparseness problem is alleviated, while the syntactic information is captured by word class. The motivation and the formulation of the class-based ngram model are similar to those of C-MEMM. Therefore, this section compares the performances of C-MEMM with those of the class-based ngram model.

First, this section compares the performance of the hard-class based MEMM with that of the class-based ngram model. The traditional bigram model is taken as the baseline model. Several class-based ngram models are built up according to the word class pattern of each internal layer of TongyiciCilin. The experimental results are presented in Table 9:

Table 9. Comparison between Hard-class based MEMM and Class-based Ngram Model

	No cluster	Clusters of 1st layer	Clusters of 2nd layer	Clusters of 3rd layer
Baseline	9.15%	---	---	---
C-Ngram	---	8.25%	7.74%	**7.37%**
C-MEMM-1	---	6.10%	**5.84%**	5.85%
C-MEMM-2	---	5.73%	5.46%	**5.28%**

From Table 9, the class-based ngram models achieve lower error rates than the baseline model, showing a more powerful predicative capability. What's more, the error rates of the class-based ngram models decrease from the 1st layer to the 3rd layer, proving that the improvement of the class-based ngram model is due to the exploitation of the increasing syntactic and semantic information of word class. However, the class-based ngram models underperformed the hard-class based MEMM models. The latter can not only make use of the syntactic and semantic information of word classes but also exploit the pinyin constraints from the input pinyin sequences.

In the following, the performance of the soft-class based MEMM with that of the class-based ngram model is compared. The POS ngram is constructed and interpolated with the traditional word ngram model. The experimental results are presented in Table 10:

Table 10. Comparison between Soft-class based MEMM and Class-based Ngram Model

	Baseline	C-Ngram	C-MEMM-1	C-MEMM-2
Error rate	8.37%	7.89%	6.00%	5.82%

The experimental results are similar to those found in Table 9. The class-based ngram models outperform the traditional ngram model by exploitation of the syntactic and semantic information in word class. However, they underperformed the soft-class based MEMM models because the latter could also make use of the pinyin constraints from pinyin sequence.

In conclusion, both the C-MEMM model and the class-based ngram model can make good use of the syntactic and semantic information of word class so as to improve the performance in the Pinyin-to-Character Conversion task; however, the former outperforms the latter by additionally exploiting the pinyin constraints from the pinyin sequence.

5. Related Works

To the best of our knowledge, there is no literature that proposes a class expansion to the MEMM model. John Lafferty [Lafferty and Suhm 1996] proposes a cluster expansion to the GIS algorithm so as to train the ME language model efficiently. However, as Lafferty admits, the technique is of little use in computing the exact ME solution. Joshua Goodman [Goodman 2001] proposes a speedup technique for the ME training process in language modeling. He decomposes the traditional ngram model into several class-based ngram models and applies the ME principle in each sub-model. There are significant differences between the Goodman 's work and this paper's. First of all, C-MEMM aims to solve the sequence label problem instead of the sequence ranking problem as language model does. Second, this paper deduces the probability function of C-MEMM based on the conditional probability of the *whole* sequence, whereas Goodman gets the probability function based on the decomposition of the *local* ngram probability. Third, this paper applies C-MEMM to the Pinyin-to-Character Conversion task in order to improve the application performance; however, Goodman is used to speed up the training process of the ME model. Moreover, both the case of hard class and soft class are discussed in this paper. In contrast, Goodman's technique is built up only in the case of a hard class.

6. Conclusions

This paper aims to improve the performance of the Pinyin-to-Character Conversion system by exploitation of the pinyin constraints from the pinyin sequence. The MEMM framework is used to describe both the pinyin constraint and the character constraint. The Class-based Maximum Entropy Markov Model (C-MEMM) is proposed to solve the efficiency problem of MEMM in the Pinyin-to-Character Conversion task. The probability functions of C-MEMM are strictly deduced and well formulized by the Bayes rule and the Markov property. Both the case of hard class and soft class are well discussed. From the experimental results, the conclusions can be drawn as follows:

- Compared with the traditional ngram model, C-MEMM improves the performance of the Pinyin-to-Character Conversion system effectively by exploitation of the pinyin constraints from the input pinyin sequences.
- C-MEMM can make good use of the syntactic and semantic information in word class and attain further improvement.
- The soft-class based MEMM outperforms the hard-class based MEMM by exploitation of more comprehensive properties of word.

Acknowledgements

This investigation is supported by the key project of the National Natural Science Foundation of China (grant No.60435020), the project of the National Natural Science Foundation of China (grant No.60673037), the project of the High Technology Research and Development Program of China (grant No.2006AA01Z197), the project of MOE-MS Key Laboratory of

Natural Language Processing and Speech in China (grant No.01307620).

Especially, the authors thank the anonymous reviewers for their valuable suggestions and comments.

References

Berger, A, S. D. Pietra, and V. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics,* 1996, 22(1), pp. 39-71.

Brown, P. F., V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language", *Computational Linguistics*, 1992, 18(4), pp. 467-479.

Chen, Y., "Chinese Language Processing", *Shang Hai education publishing company*. 1997

Chen, L. Z. and T. Y. Huang, "A Novel Word Clustering Algorithm And Vari-gram Language Model", *Journal of Computer Sciences*, 1999, 22(9), pp. 942-948.

Chen, Z, and K. F. Lee, "A New Statistical Approach To Chinese Pinyin Input", *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL2000)*, Hong Kong, 3-6 October 2000.

Darroch, J. N. and D. Ratcliff, "Generalized Iterative Scaling for Log-linear Models". *Annals of Mathematical Statistics*, 1972, 43, pp. 1470-1480.

Emerson, T. "The Second International Chinese Word Segmentation Bakeoff". *In Proceedings of The Fourth Sighan Workshop on Chinese Language Processing*, 2005, pp. 123-133.

Gao, J. F., J. Goodman, and J. B. Miao, "The Use of Clustering Techniques for Language Modeling - Application to Asian languages". *International Journal of Computational Linguistics and Chinese Language Processing,* 6(1), pp. 27-60. 2001.

Gao, J. F, J. Goodman, M. J. Li, K. F. Lee, "Toward a unified approach to statistical language modeling for Chinese", *ACM Transactions on Asian Language Information Processing*, 2002, 1(1), pp. 3-33.

Gao, J. F, H. Yu and W. Yuan, "Minimum Sample Risk Methods for Language Modeling". *In Human Language Technology Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Oct 6-8, Vancouver, Canada, 2005.

Goodman, J., "Classes for fast maximum entropy training". *Proceedings of the IEEE International Conference on Acoustics, Speach and Signal Processing (ICASSP-2001)*, IEEE press, 2001.

Hsu, W. L. and K. J. Chen, "The Semantic Analysis in GOING - An Intelligent Chinese Input System", *Proceedings of the Second Joint Conference of Computational Linguistics*, Shia men, 1993, pp. 338-343.

Hsu, W. L., "Chinese parsing in a phoneme-to-character conversion system based on semantic pattern matching", *International Journal on Computer Processing of Chinese and Oriental Languages,* 1995, volume 40, pp. 227-236.

ISO, "Information and documentation - Romanization of Chinese", ISO 7098:1991

Jeffreys, H, "Theory of Probability". *Clarendon Press*, Oxford, second Edition. 1948.

Jelinek, F. and R. L. Mercer, "Interpolated estimation of markov source parameters from sparse data". *In Pattern Recognition in Practice,* 1980, pp. 381-397.

Katz, S. M. "Estimation of probabilities from sparse data for the language model component of a speech recognizer", *IEEE Transactions on Acoustics, Speeech and Signal Processing,* 1987, 35(3), pp. 400-401.

Kuo, J. J., "Phonetic-input-to-character conversion system for Chinese using syntactic connection table and semantic distance", *Computer Processing and Oriental Languages,* 1995, 10(2), pp. 195-210.

Lafferty, J. and B. Suhm. "Cluster expansions and iterative scaling for maximum entropy language models". *Maximum Entropy and Bayesian Methods*, K. Hanson and R. Silver, eds., Kluwer Academic Publishers, 1996.

Li, H., "Word Clustering and Disambiguation Based on Co-occurrence Data", *Proceedings of COLING-ACL98*. Montreal,Canada. 10-14 August, 1998.

Liu, T. et al., "TongYiCiCiLin (Extension Version)". 2005. Http://www.ir-lab.org

McCallum, A., D. Freitag and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation", *Proceedings of ICML2000*, Stanford, CA, USA, 2000, pp. 591-598.

Mei, J. J, Y. M. Zhu, Y. Q. Kao, H. X. Yan. "TongYiCiCiLin". Shanghai: *Shanghai Lexicographical Publishing House*. 1983.

Myung I. J, "Tutorial on Maximum Likelihood Estimation", *Journal of Mathematical Psychology,* 2003, 47, pp. 90-100.

Pietra, S. D, V. D. Pietra and J. Lafferty. "Inducing Features of Random Fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1997. 19(4), pp. 380-393.

Tsai, J. L. and W. L. Hsu, "Applying an NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem", *Proceedings of COLING02*, Taipei, 2002.

Tsai, J. L., T. J. Chiang and W. L. Hsu, "Applying Meaningful Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem", *Proceedings of ROCLING04,* 2004.

Tsai, J. L., "Applying a Mix Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem", *In 2th International Joint Conference on Natural Language Processing (IJCNLP 2005)*, Jeju, Korea, Oct 11-13, 2005.

Tsai, J. L., "Using Word Support Model to Improve Chinese Input System", *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING -ACL06)*, Sydney, Australia, 17-21 July 2006.

Wang, X. L., "Chinese Input system by Pinyin Sentence: Insun", *Journal of Chinese Information Processing*, 1993, 7(2), pp. 45-54.

Wang, X. L, Q. C Chen and D. S. YEUNG, "Mining pinyin-to-character conversion rules from large-scale corpus:a rough set approach", *IEEE Transaction on Systems Man and Cybernetics*, Part B: Cybernetics, 2004, 34(2), pp. 834-844.

Wang, X. L., D. S . Yeung, J. N. K. Liu, R. W. P. Luk and X. Wang, "A Hybrid Language Model Based on Statistics and Linguistic Rules", *International Journal of Pattern Recognition and Artificial Intelligence*, 2005, 19(1), pp. 109-128.

Wang, Y. M., "The Three Principles of Computer Chinese Character Keyboard Design", *Chinese Journal of Computers*, 2005, 28(5), pp. 870-881.

Wu, J., "Implementation and Application of Statistical Language Model in Mandarin Speech Recognition", *master dissertation of the Tsinghua University,* 2000.

Xiao, J. H, B. Q Liu and X. L. Wang, 2005a, "Principles of Non-stationary Hidden Markov Model and its Applications on Sequence Labeling Task". *In Proceedings of the 2th International Joint Conference on Natural Language Processing (IJCNLP 2005)*, Jeju, Korea, Oct 11-13, 2005a.

Xiao, J. H., B. Q. Liu and X. L. Wang, 2005b, "A Similarity-based Approach to Data Sparse Problem of Chinese Language Modeling", *In Proceedings of the 4th Mexico International Conference on Artificial Intelligent (MICAI2005)*. pp. 761-769, Best Student Paper Award. Mexico, November 14-18, 2005b.

Xiao, J. H., B. Q. Liu and X. L. Wang, "A Similarity-Based Smoothing Algorithm for Chinese Language Modeling and its Application on Pinyin-to-Character Conversion", *High Technique Letters*, 2006, 16(2), pp. 127-132.

Xu, Z. M., X. L. Wang and S. X. Jiang, "A Sentence-Level Chinese Character Input Method", *High Technique Letters*, 2000, (1), pp. 51-56.

Yu, S. W., H. M. Duan, B. Swen and B. B. Chang, "Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation". *Journal of Chinese Language and Computing*, 2003, 13(2), pp. 121-158.

Zhang, R. Q, Z. Y. Wang and J. P Zhang, "Chinese Pinyin-to-Text Translation technique with Error Correction used for Continuous Speech Recognition", *Journal of Tsinghua University*, 1997, 37(10), pp. 9-12.

Zhang, R. Q., Z. Y. Wang and D. J. Lu, "Zero-Probablities of Language Model in Translations of Chinese Spellings to Characters", *Acta Electornica Sinica,* 1998, 26(8), pp. 43-46.

Zhou G. D. and K. T. Lua. "Word Association and MI-Trigger-based Language Modeling". *Proceedings of COLING-ACL98*. Montreal,Canada. 10-14 August, 1998.

Zipf, G. K., "Psycho-Biology of Languages", *The MIT Press*. 1935.