PAPER Special Section on Robust Speech Processing in Realistic Environments

Automatic Language Identification with Discriminative Language Characterization Based on SVM

Hongbin SUO[†*a)], ***Student Member***, **Ming LI**[†], **Ping LU**[†], ***and*** **Yonghong YAN**[†*b)], ***Nonmembers***

SUMMARY Robust automatic language identification (LID) is the task of identifying the language from a short utterance spoken by an unknown speaker. The mainstream approaches include parallel phone recognition language modeling (PPRLM), support vector machine (SVM) and the general Gaussian mixture models (GMMs). These systems map the cepstral features of spoken utterances into high level scores by classifiers. In this paper, in order to increase the dimension of the score vector and alleviate the inter-speaker variability within the same language, multiple data groups based on supervised speaker clustering are employed to generate the discriminative language characterization score vectors (DLCSV). The back-end SVM classifiers are used to model the probability distribution of each target language in the DLCSV space. Finally, the output scores of back-end classifiers are calibrated by a pair-wise posterior probability estimation (PPPE) algorithm. The proposed language identification frameworks are evaluated on 2003 NIST Language Recognition Evaluation (LRE) databases and the experiments show that the system described in this paper produces comparable results to the existing systems. Especially, the SVM framework achieves an equal error rate (EER) of 4.0% in the 30-second task and outperforms the state-of-art systems by more than 30% relative error reduction. Besides, the performances of proposed PPRLM and GMMs algorithms achieve an EER of 5.1% and 5.0% respectively.

key words: *language identification, supervised speaker clustering, support vector machine, discriminative language characterization score vector, pair-wise posterior probability estimation*

1. Introduction

Automatic spoken language identification without using deep knowledge of those languages is a challenging task. The variability of one spoken utterance can be incurred by the content, speakers and environment. Normally the training corpus and test corpus consist of unconstrained utterances from different speakers. Therefore, the core issue is how to extract the language differences regardless of content, speaker, and environment information [1], [2]. The clues that human use to identify languages are studied in [3], [4]. The sources of information used to discriminate one language from the others include phonetics, phonology, morphology, syntax and prosody. At present, most reported automatic LID systems take advantage of one or more of these language traits in the identification task.

A number of researchers have used phone recognizers as front-end for language identification in [5]–[12]. The most successful approach to LID uses phone recognizers of several languages in parallel. In [7], it is shown that even with one language phone recognizer, an LID system can be built. It is named phone recognition language modeling (PRLM). However, the analysis in [4] also indicates that the performance of a system can be considerably improved in proportion to the number of front-end phone recognizers. Recently, a set of phone recognizers are used to transcribe the input speech into phoneme lattices [13], [14] which are later scored by n-gram language models.

Several recent approaches using support vector machine (SVM) and Gaussian mixture model (GMM) have attracted much attention as an alternative solution. Due to the introduction of shifted-delta-cepstral (SDC) acoustic features, promising results using SDC are reported [15], [16]. This approach is further improved by using discriminative Maximum Mutual Information (MMI) training for acoustic modeling in GMMs [17]. In order to be speaker independent, a set of speaker dependent anchor GMM models are trained on SDC features for every speaker in every language, and back-end discriminative SVM classifiers are adopted to identify the spoken language based on the GMM outputs [18].

Vector space modeling approach has been successfully applied to spoken language identification. The acoustic characteristics of spoken language are collected into acoustics segment models (ASMs) [19], and each spoken utterance is converted into a feature vector with its attributes representing the statistics of the acoustics units, thus a discriminative classifier is built in this score vector space to identify the target language. The main object of these improvements is to derive the discriminative high level feature vectors in LID tasks while restraining the disturbance caused by the variability of speakers or channels in realistic system. Results in anchor GMM system [18] show that it is capable to achieve robust speaker independent language identification through compensation for intra-language and inter-speaker variability.

However, for every test speech segment, scoring on these entire anchor-GMM language models is computationally expensive, also sufficient training data for each anchor speaker can hardly be guaranteed in practical application. Moreover, the identity of a target language is not sufficiently described by the score vectors which are generated by the following language models in conventional PPRLM systems. To compensate these insufficiencies, it is a natural extension that multiple groups with similar speakers in one

Manuscript received July 4, 2007.
Manuscript revised September 19, 2007.

[†]The authors are with the ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, China.
[*]Correspondence author
a) E-mail: hsuo@hccl.ioa.ac.cn
b) E-mail: yyan@hccl.ioa.ac.cn

DOI: 10.1093/ietisy/e91–d.3.567

language are used to build the multiple target phonotactic language models or acoustic models. For example, the training data set can be divided by genders. This approach can also be applied to SVM system. Each target score in generalized linear discriminant sequence (GLDS) kernel [16] can be calculated simply by an inner product, and multiple classifiers are created for different speaker groups in extremely large database. This is followed by fusing multiple kernels with different weights in discriminative language characterization score vector space. Thus a new scoring function is generated with less speaker dependence. In this paper, the hierarchical clustering (HC) algorithm and K-means clustering algorithm are used together to extract more information from the available training data.

At the back-end of the LID systems, SVM based classifiers have demonstrated superior performance over generative language modeling framework in [19]–[21]. SVM as a discriminative tool maps input cepstral feature vector into high-dimensional space and then separates classes with maximum margin hyperplane. In addition to its discriminative nature, its training criteria also balance the reduction of errors on training data and the generalization on unseen data. This makes it perform well on small quantities of data and suited for handling high dimensional problem. In this paper, a back-end radial basis function (RBF) kernel [22] SVM classifier is carried out to discriminate target languages based on the probability distribution in discriminative language characterization score vector space. The choice of radial basis function kernel is based on its nonlinear mapping function and relatively small amount of parameters to tune. Furthermore, the linear kernel is a special case of RBF and the sigmoid kernel behaves like radial basis function for certain parameters [23]. Note that training data of this back-end SVM classifier comes from development data rather than the data used for training front-end GLDS classifiers, and cross validation is employed to select kernel parameters and prevent over-fitting problem. For testing, once the discriminative language characterization score vectors of a test utterance are generated, back-end SVM classifier can estimate the posterior probability of each target language, which is used to calibrate final outputs.

In this paper, there are many unfamiliar abbreviations. For avoiding confusing, these abbreviations are clearly explained again as follows:

PRLM	phone recognition language modeling
PPRLM	parallel phone recognition language modeling
SDC	shifted-delta-cepstral
LRE	language recognition evaluation
GLR	generalized likelihood ratio
GLDS	generalized linear discriminant sequence
DLCSV	discriminative language characterization score vectors
ASM	acoustics segment model
LLR	log-likelihood ratio

The remainder of this paper is organized as follows. In Sect. 2, a speech corpus used for this study is introduced. A supervised clustering algorithm is described in Sect. 3. How the DLCSV space is used to improve conventional LID systems is briefed in Sect. 4. In Sect. 5, the SVM classifier with RBF kernel is detailed. Besides, a score calibration method and a probability estimation algorithm are explained in this section. Experiments and results of the proposed method are given in Sect. 6. Finally, a brief summary is discussed in Sect. 7.

2. Speech Corpus

In phone recognizer framework, the Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [24] is used. It contains 90 speech messages in each of the following 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Each message is spoken by a unique speaker and comprises responses to 10 prompts. Besides, phonetically transcribed training data is available for six of the OGI Languages (English, German, Hindi, Japanese, Mandarin and Spanish). Otherwise, the labeled Hong Kong University of Science and Technology (HKUST) Mandarin Telephone Speech Part 1 [25] is used to accurately train an acoustic model for another Mandarin phone recognizer. A telephone speech database in common use for back-end language modeling is the Linguistic Data Consortium's CALLFRIEND corpus. The corpus comprises two-speaker, unprompted, conversational speech messages between friends. Hundred North-American long-distance telephone conversations are recorded in each of twelve languages (the same as 11 languages as OGI-TS plus Arabic). There are three sets in this corpus including training, development and test set, each set consists of 20 two-sided conversations from each language, approximately 30 minutes long.

Moreover, development data which can be used to tune the parameters of back-end classifiers is obtained from the 1996 NIST LRE development and evaluation sets. And, experiments are performed on the 2003 NIST LRE [26] 30 s test set. Here, 960 true trials and 10560 false trials are generated from the primary evaluation task. Thus, the data comprises 80 test segments, for each of the 12 target languages (the same as the CALLFRIEND corpus). All of the training, development and evaluation data is in standard 8-bit 8 kHz mu-law fromat from digital telephone channel.

3. Supervised Speaker Clustering

This section mainly introduces how the training data of each language is divided into several subgroups by individuality. Hierarchical clustering (HC) algorithm [27] followed by K-means clustering is proposed in this section.

3.1 Hierarchical Clustering

K is denoted as the number of speakers in one language corpus, L is the number of target languages and N is set with

the number of target subgroups in one language corpus.

Consider a collection of total speech segments of one language $X = \{x_1, x_2, \ldots, x_K\}$ and each x_i represents a sequence of spectral feature vectors. Here, each speaker is corresponding to one speech segment. The algorithm can be described as follows:

1. Initialize the number of clusters $c : c \leftarrow K$
2. Compute pair-wise distances between each cluster
3. Find the nearest pair in the c clusters: x_i and x_j
4. Merge x_i and x_j as a new cluster
5. Update distances of cluster to new cluster
6. Update the hypothesized number $c : c \leftarrow (c-1)$
7. Iterate steps 3-6 until $c = N$

Here, generalized likelihood ratio (GLR) distance is chosen as the pair-wise distances between two clusters.

3.2 Generalized Likelihood Ratio Distance

Consider two speech segments x and y from different speakers. Assume that $L(x;\mu_x,\sum_x)$ and $L(y;\mu_y,\sum_y)$ denote the likelihood of x for single Gaussian model $N(\mu_x,\sum_x)$ and that of y for single Gaussian model $N(\mu_y,\sum_y)$ respectively. The likelihood of attribute two segments from the same speaker is given by $L(z;\mu_z,\sum_z)$, where z is the union of segments x and y. The generalized likelihood ratio [28] is defined by:

$$\begin{aligned} GLR(x,y) &= -\log\left[\frac{L(z;\mu_z,\sum_z)}{L(x;\mu_x,\sum_x)\times L(y;\mu_y,\sum_y)}\right] \\ &= -\log\left[\frac{|\sum_x|^{\alpha}|\sum_y|^{1-\alpha}}{|W|}\right]^{\frac{M}{2}} \\ &\quad -\log\left[1+\frac{M_xM_y}{M^2}(\mu_x-\mu_y)^T W^{-1}(\mu_x-\mu_y)\right]^{\frac{M}{2}} \end{aligned} \tag{1}$$

where M_x and M_y are the number of frames from speech segments x and y respectively, $M = M_x + M_y$, $\alpha = M_x/M$, and W is their frequency weighted average $W = (M_x\sum_x + M_y\sum_y)/M$. The generalized likelihood ratio distance has been found useful in the hierarchical clustering framework. More details can be found in [29].

4. Proposed LID Frameworks

In preceding section, the supervised speaker clustering algorithm is explained firstly. The following sections mainly introduce some improved frameworks based on clustered speaker groups and discriminative language characterization score vectors.

4.1 Parallel Phone Recognizer with Language Model Groups

Parallel phone recognizer with language model groups system is composed of four parts [30]: feature extractor, language dependent phone recognizers, score generators and back-end classifier. The general system architecture for language identification task is given in Fig. 1, where PRi and SGi are language-dependent phone recognizer and score generator for language i. Acoustic scores (likelihood) are generated by one pass Forward-Backward decoding algorithm. And, phonotactic scores are generated by the following language models in score generators. Usually, the number of phonotactic scores is equal to the number of target languages. Final score vector which is composed of the two types of scores is sent to back-end classifier for identification.

Figure 2 shows a Mandarin score generator. In the scoring framework, the training set of each target language is divided into multiple groups which are used to build corresponding language models. Thus, the dimension of score vector is extended to high level. The total number of language models is $N_{total} = L \times N$, as defined in Sect. 3.1, L is the number of target languages and N is the number of target subgroups in one language corpus. So, taking no count of the acoustic scores, the dimension of discriminative language characterization score vectors in PPRLM system is $N_{DLCSV} = L \times N \times P$, where P is denoted as the number of phone recognizers in parallel. Considering the amount of training data for language model building, N is limited to a small number.

4.2 Support Vector Machine Groups

In this section, a proposed method focuses on how to make SVM-SDC technique more effective. Standard SVM-SDC

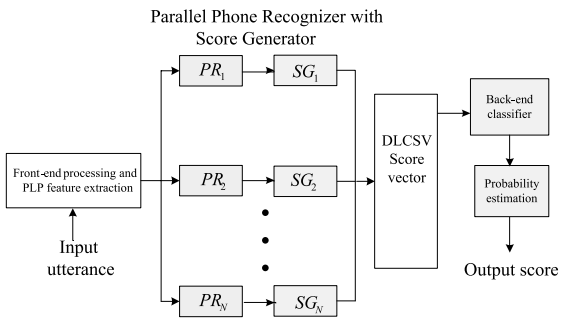


Fig. 1 Structure of the proposed PPRLM system.

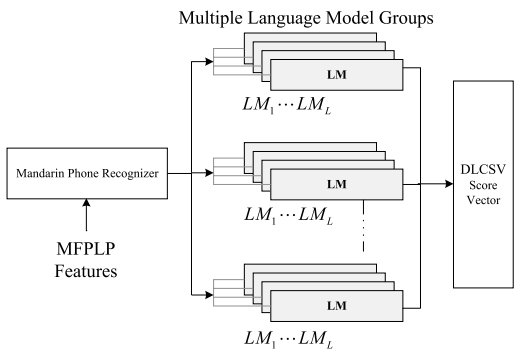


Fig. 2 Structure of the Mandarin score generator.

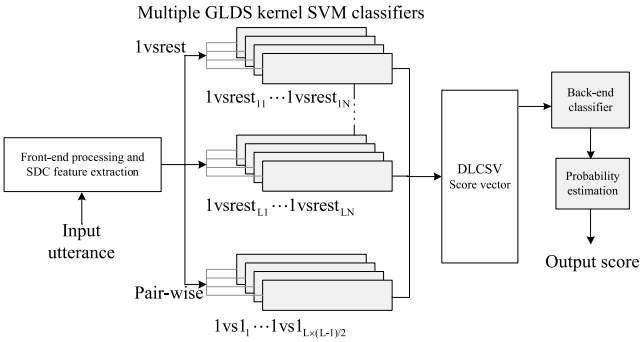


Fig. 3 Structure of the proposed SVM system.

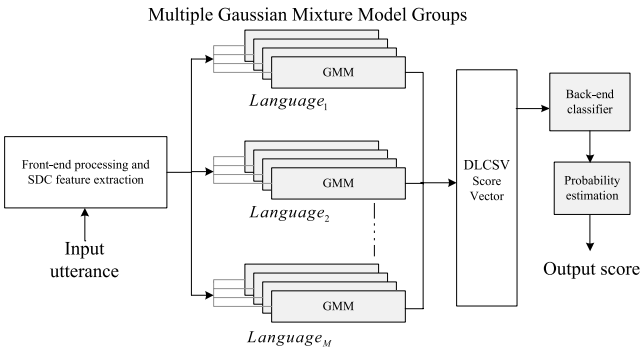


Fig. 4 Structure of the proposed GMM system.

framework [16] is improved in this paper by employing multiple GLDS kernel classifiers to convert each speech utterance into a score vector, which is the combination of all classifiers' output scores, and represents the extent of matching between each input utterance and all the language specific constraints on the margins. In the proposed framework, multiple speaker group dependent SVM classifiers are trained to map the expanded cepstral features of speech segments into discriminative language characterization score vectors. As demonstrated in Fig. 3, both pair-wise and parallel one-versus-the-rest discriminative classifiers are trained by the generalized linear discriminant sequence approach. The total numbers of GLDS kernel classifiers are $N_{total} = L \times (L-1)/2 + L \times N$. Thereby, each target language data of CALLFRIEND corpus is divided into N subgroups by speaker clustering algorithm, each of which represents a set of different speakers speaking the same language. By training each group-based one-versus-the-rest or pair-wise classifier separately, a problem regarding memory limitation is fixed. Moreover, experiments show that significant performance improvement can be obtained by compensating these distortions in the domain of discriminative language characterization score vectors which result from the inter-speaker variability presented by different speaker groups within the same language.

SVMTorch [22] is used to train all these classifiers with above mentioned kernel function. After discriminative training, multiple classifiers are generated by combining these multiple classifiers' output scores together. Each input speech utterance can be mapped into a single DLCSV. In the framework shown in Fig. 3, multiple speaker group dependent SVM classifiers with generalized linear discriminant sequence kernel are trained to map the expanded cepstral features of speech segments into discriminative language characterization score vectors.

4.3 Gaussian Mixture Modeling Based Speaker Groups

As shown in Fig. 4, the speaker groups GMM system is similar to an anchor GMM system. As mentioned in Sect. 1, the anchor GMM is built by modeling for each speaker data of each language corpus. The anchors are a predetermined set of speakers that is non-intersecting with the set of target speakers in the test utterances [18]. In contrast to anchors, the speaker groups are generated by clustering. In other words, each language probability distribution in DLCSV space is described through multiple Gaussian Mixture Models. The total number of GMMs are $N_{total} = L \times N$. Obviously, along with the increase of Gaussian mixtures, the computational complexity will also increase more greatly.

5. Proposed Back-End Classifier

In the LID frameworks, after discriminative language characterization score vectors of test utterances are generated, back-end SVM classifier estimates the posterior probability of each target language, which is used to calibrate the final outputs. This section mainly introduces a proposed back-end classifier algorithm and a score calibration algorithm.

5.1 RBF Support Vector Machine

An SVM is a two-class classifier constructed from sums of a kernel function $K(.,.)$

$$f(x) = \sum_{i=1}^{n} \alpha_i t_i K(x, x_i) + d, \quad \text{subject to } \alpha_i > 0 \text{ and } \sum_{i=1}^{n} \alpha_i t_i = 0 \tag{2}$$

where n is the number of support vectors, t_i is the ideal outputs, α_i is the weight for the support vectors x_i. A back-end radial basis function (RBF) [22] kernel is carried out to discriminate target languages. RBF kernel is defined as follows:

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0 \tag{3}$$

where γ is the kernel parameter estimated from the training data.

5.2 Score Calibration

The topic of calibrating confidence scores in the field of multiple-hypothesis language recognition has been studied in [31], and a detailed analysis of the information flow and

the amount of information delivered to users through a language recognition system has been performed. The posterior probability of each of M hypotheses is estimated and a maximum-a-posteriori (MAP) decision is made. In [16], log-likelihood ratio (LLR) normalization which has been proved to be useful is adopted as a simple back-end process. In the normalization, suppose $\overrightarrow{S} = [S_1, S_2, \ldots S_L]^t$ is the vector of L relative log-likelihoods from L target languages for a particular message, and the posterior probabilities for original hypotheses can be denoted as:

$$P_i = \pi_i e^{S_i} \Big/ \left(\sum_{j=1}^{L} \pi_j e^{S_j} \right), i = 1, 2, \ldots, L \tag{4}$$

where $[\pi_1, \ldots, \pi_L]$ denotes the prior. Considering a flat prior, new log-likelihood ratio normalized score S_i' is denoted as

$$S_i' = S_i - \log \left(\frac{1}{M-1} \sum_{j \neq i} e^{S_j} \right) \tag{5}$$

However, the output scores of back-end RBF SVM are not log-likelihood values, thus linear discriminant analysis (LDA) and diagonal covariance Gaussian models are used to calculate the log-likelihoods for each target language [32], and improvement has been achieved in detection performance [16].

In this paper, we proposed an alternative approach [23] to estimate the posterior probabilities. Given L classes of data, the goal is to estimate $p_i = p(y = i|x), i = 1, \ldots L$. In a pair-wise framework, firstly pair-wise class probabilities are estimated as

$$r_{ij} \approx p(y = i | y = i \text{ or } j, x) \approx 1 \Big/ (1 + e^{A\hat{f}+B}) \tag{6}$$

where A and B are estimated by minimizing the negative log-likelihood function using known training data and their decision values $\hat{f}$. Then, posterior probability p_i can be obtained by optimizing the following problem:

$$\begin{aligned} &\min \frac{1}{2} \sum_{i=1}^{L} \sum_{j \neq i} (r_{ji} p_j - r_{ij} p_i)^2, \\ &\text{subject to } \sum_{i=1}^{L} p_i = 1 \text{ and } p_i > 0 \end{aligned} \tag{7}$$

Therefore, the estimated posterior probabilities are applicable to performance evaluation. The probability tools of LIBSVM [22] are used in our approach. Experiments in next section show that this multi-class pair-wise posterior probability estimation algorithm is superior to commonly-used log-likelihood ratio normalization method.

6. Experiments and Results

The performance of a detection system is characterized by its miss and false alarm probabilities. The primary evaluation metric is based upon 2003 NIST language recognition evaluation [26]. The task of this evaluation is to detect the presence of a hypothesized target language, given a segment of conversational speech over the telephone. The target language will be one of the following twelve languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Submitted scores are given in the form of equal error rates (EER). EER is the point where miss probability and false alarm probability are equal. Experiments of the proposed application are explained in following sections.

6.1 Performance of Phone Recognizer Systems

In feature extractors of phone recognizer systems, speech data is parameterized every 25 ms with 15 ms overlap between contiguous frames. For each frame a feature vector with 39 dimensions is calculated: 13 Mel Frequency Perceptual Linear Predictive (MFPLP) [33], [34] coefficients, 13 delta cepstral coefficients, 13 double delta cepstral coefficients. All the feature vectors are processed by cepstral mean subtraction (CMS) method.

A Mandarin phone recognizer is built from HKUST Telephone data in a PRLM system. There are 68 monophones and a three-state left-to-right Hidden Markov Models (HMM) is used for each tri-phone in each language, with which acoustic model is described in more detail. But, PPRLM system is mainly composed of six phone recognizers. Acoustic model for each phone recognizer is initialized on OGI-TS corpus and retrained on CALLFRIEND training set corpus. Since the amount of labeled data is limited, mono-phone is chosen as acoustic modeling unit.

The outputs of all recognizers are phone sequences which are used to build the following 3-gram phone language models. And, for comparing with other systems in phone recognizer frameworks, only the phonotatic scores which are log-likelihoods generated by language models scoring are composed to DLCSV for classifying.

In this paper, the pair-wise posterior probability estimation algorithm is proposed to calibrate the language scores. Besides, diagonal covariance Gaussian model followed by log-likelihood ratio normalization algorithm is evaluated for comparison. However, it is hardly describe distribution of the high dimensional DLCSV using Gaussian model. Linear discriminant analysis method is used to reduce the high dimension of score vectors [16]. In the mean time, a feed-forward neural network (NN) is used as the back-end classifier for another competent system [35].

The equal error rate performances of ten systems with phone recognizer algorithm are given in Tables 1 and 2. The main frameworks which are composed by discriminative language characterization score vectors and the followed different back-end classifiers are checked with marks. Firstly, the baseline systems are denoted as DLCSV12 and DLCSV72 for no speaker cluster in the phone recognizer framework. Then, the 12 dimensional scores of PRLM-DLCSV12 can be used to identify the target language, with no any classifier. Besides, the high dimensional scores can

Table 1 PRLM systems results on 2003 NIST 30 s tasks.

PRLM system	1	2	3	4	5
DLCSV12	√	√			
DLCSV24			√	√	√
LLR		√			
NN			√		
LDA+GM+LLR				√	
SVM+PPPE					√
EER (%)	11.9	11.7	10.0	10.7	9.7

Table 2 PPRLM systems results on 2003 NIST 30 s tasks.

PPRLM system	1	2	3	4	5
DLCSV72	√	√	√		
DLCSV144				√	√
NN	√				
LDA+GM+LLR		√		√	
SVM+PPPE			√		√
EER (%)	6.7	6.1	5.7	5.8	5.1

be generated by multiple language models with subgroups. Considering the amount of training data for language modeling, the target number of subgroups is set to 2 (female and male). Thus, the dimension of the DLCSV is 24 in PRLM framework and 144 in PPRLM framework. Secondly, NN, LLR and PPPE algorithms are evaluated respectively on LRE task for comparing to each other.

6.2 Performance of SVM Groups System

In SVM system, after front-end processing, 56 dimensional SDC features are extracted as in [16]. And, all polynomials up to degree 3 are used to expand the primary features into an expansion with dimension of 32509. The numbers of target languages L and sub-speaker groups in each language N are respectively set to 12 and 6. Here, the number of subgroups is chosen as 6 for two reasons. Firstly, considering memory limitation, pair-wise classifiers only need to load training samples from two languages rather than all of the twelve target languages, which require less memory and allow training processes to use more samples for each language. Thus, all of the remaining language data could be uploaded in one-versus-the-rest classifier training. Secondly, in DLCSV-138, 12 one-versus-the-rest classifiers are replaced by multiple speaker group based classifiers, which represent both discriminative language information and inter-speaker variability within the same language. By using back-end classifiers, this speaker group specified variability can be compensated and make system less speaker dependent. Thus, the total number of GLDS SVM classifiers N_{total} is 138.

Five types of experiments are conducted to evaluate the performance of each part of the proposed methods with check marks in Table 3. Firstly, Score vector modeling [19] approach is evaluated in systems 1-3. The procedures are detailed as follows: after one utterance is classified by multiple SVM classifiers, the generated scores are combined into a score vector which is used to train a high-level model or classifier. DLCSV12 denotes that only 12 one-versus-the-rest classifiers are used to construct the DLCSV space and the duration of each utterance is 3 minutes, while DLCSV78 uses both 12 one-versus-the-rest and 66 pair-wise classifiers to map the input speech utterance into DLCSV space. In DLCSV138 approach, the dimension of score vector is 138 which combines multiple classifiers' outputs together, including both 66 one-versus-one and group based one-versus-the-rest classifiers. The duration of speech segments used for training these 138 classifiers is 30 seconds. Secondly, pair-wise posterior probability estimation and log-likelihood ratio normalization algorithms are evaluated respectively to calibrate the output language scores. At last, SDC feature with the parameters of 7-1-3-7 is replaced by the modified 56 dimension SDC features described in [17] to enhance the capability of language discrimination.

Table 3 SVM system results on NIST 2003 30 s tasks.

SVM-SDC system	1	2	3	4	5
DLCSV12	√				
DLCSV78		√			
DLCSV138			√	√	√
LLR	√	√	√		
PPPE				√	√
SDC 7-1-3-7	√	√	√	√	
SDC 7-1-3-7 + MFCC					√
EER (%)	7.0	5.9	5.0	4.7	4.0

Table 4 GMM-SDC system results on NIST 2003 30 s task.

GMM-SDC Systems	EER (%)
Baseline	9.3
DLCSV24+LDA+GM	8.7
DLCSV24+LDA+GM+LLR	8.3
DLCSV24+SVM+PPPE	7.8
DLCSV72+LDA+GM+LLR	7.1
DLCSV72+SVM+PPPE	6.5
DLCSV144+LDA+GM+LLR	6.6
DLCSV144+SVM+PPPE	5.9
DLCSV240+SVM+PPPE	5.1
DLCSV480+SVM+PPPE	5.0

6.3 Performance of GMM Groups Systems

Since acoustic level features are particularly analyzed and proved to be useful for discriminating one language from another. The modified 56 dimensional SDC features are also used in GMM systems. The baseline system is built based on maximum likelihood (ML) algorithm, and each Gaussian mixture model corresponds to one target language. The following systems shown in Table 4 are based on sub-speaker groups. As mentioned in Sect. 3.1, the number of supervised clusters for one language N is chosen as 2, 6, 12, 20 and 40, respectively. Thus, the dimension of DLCSV which are generated by front-end GMM scoring is 24, 72, 144, 240 and 480, respectively. When N is equal to 2, the corpus is almost divided by female and male. Thus, the system is based on a gender dependent Gaussian mixture models. The number of Gaussian mixture is set to 512 for comparing with anchor GMM. In addition, pair-wise posterior probability

Table 5 Comparison with state-of-art systems.

LID Systems	EER (%)
Baseline	9.3
PPRLM-MIT [32]	6.6
PPRLM-LIMSI [13] (string)	6.8
PPRLM-LIMSI [13] (lattice)	2.7
Proposed PPRLM system 5	5.1
GLDS SVM	6.1
Proposed SVM-SDC system 5	4.0
Anchor GMM [18]	4.8
Proposed GMM	5.0

Table 6 The computational cost of proposed systems.

LID Systems	Real Time (RT)
PPRLM system 1	0.743
PPRLM system 2	0.728
PPRLM system 3	0.716
PPRLM system 5	0.739
SVM system 1	1.12×10^{-3}
SVM system 3	1.19×10^{-3}
SVM system 5	1.96×10^{-3}
GMM Baseline	8.6×10^{-3}
Proposed GMM	0.171
Anchor GMM	0.634

estimation (PPPE) and log-likelihood ratio (LLR) normalization approaches are also adopted in experiments of Gaussian mixture modeling system.

6.4 Discussion

The experiment results of DLSCV systems show that discriminative score vector modeling method improves system performance in most cases. As mentioned above, the main reason is that multiple discriminative classifiers based on hierarchical clustered speaker groups are employed to map the speech utterance into discriminative language characterization score vector space, which not only represents enhanced language information but also compensates for intra-language and inter-speaker variability. Moreover, by using back-end classifiers, this speaker group specific variability can be compensated sufficiently and make system less speaker dependent. Furthermore, as shown in Tables 1-4, the proposed PPPE method adopted in improved systems is comparable to the common employed LLR approach. Because the output scores of back-end classifiers are not real log-likelihood values, this alternative language score calibration method performs better. And, SDC feature concatenated with MFCC coefficients achieves significant improvement as demonstrated in SVM system. Obviously, in GMM systems, the performance is improved gradually along with the increase of subgroups. The comparison of performance with other systems is shown in Table 5. After comparing the results of phone recognizer systems, using lattice information to build language models and scoring by lattice can improve the performance notably.

Computational cost of the proposed algorithm is low, compared with the conventional systems. The main reasons can be explained as follows. Firstly, the improved back-end SVM classification with PPPE algorithm requires a low computational cost. Secondly, the increment of computational cost is focused on generating the discriminative language characterization score vectors. Thus, in PPRLM system, the time cost of language model scoring is much lower than phone recognizing. In SVM system, the main computational effort is spent on expanding features from low dimension to high dimension. Whereas, the computational cost of GMMs system is raised along with the number of Gaussian mixture models according to speaker subgroups. Table 6 shows the computational cost of the most systems in this paper. The evaluations are carried out on a machine with 3.4 G Hz Intel Pentium CPU and 1 G Byte memory.

7. Conclusions and Future Work

In this paper, a novel approach using a supervised hierarchical algorithm to initialize the speaker groups for further K-means clustering is introduced in detail. The progressive use of the groups' training data for building language models, support vector machine classifiers and Gaussian mixture models are exploited to map the speech utterance into discriminative language characterization score vector space efficiently. This feature set represents enhanced language information, and at the same time compensates the disturbances caused by intra-language and inter-speaker variability. The new approach is applied to enhance mainstream systems including PPRLM, SVM and GMMs systems. Experiment results on 2003 NIST language evaluation task demonstrate that significant improvement is achieved by mapping speech utterance into this DLCSV feature space. Furthermore, the traditional back-end classification using Gaussian classifier with language log-likelihood ratio normalization is replaced by new methods. Both back-end RBF kernel support vector machine classifier and pair-wise posterior probability estimation methods are proposed and investigated to further improve the performance.

Recently, one common practice in large vocabulary continuous speech recognition (LVCSR) is to exploit rich information such as lattice at the end of first pass decoding. In LID task, this approach of using lattices instead of phone sequences has been reported with improved performance. Furthermore, it is generally believed that phonotactic feature and acoustic cepstral feature provide complementary cues to each other. The fusion of multiple information sources has been proven to be effective in recent studies. These directions will be exploited in our future work.

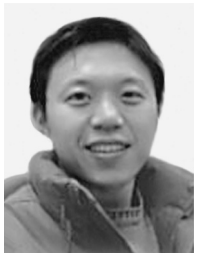
Acknowledgments

This work is partially supported by the Ministry of Science and Technology of the People's Republic of China (973 Program, 2004CB318106), National Natural Science Foundation of China (10574140, 60535030), The National High Technology Research and Development Program of China (863 Program, 2006AA010102, 2006AA01Z195). A pre-

liminary version of this work was presented in 2007 InterSpeech.

References

[1] K.P. Li, "Automatic language identification using syllabic spectral features," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.297–300, April 1994.
[2] T. Nagarajan and H.A. Murthy, "Language identification using acoustic log-likelihoods of syllable-like units," Speech Commun., vol.48, no.8, pp.913–926, Jan. 2006.
[3] Y.K. Muthusamy, N. Jain, and R.A. Cole, "Perceptual benchmarks for automatic language identification," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.333–336, April 1994.
[4] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech Audio Process., vol.4, no.1, pp.31–44, Jan. 1996.
[5] L.F. Lamel and J.L. Gauvain, "Language identification using phone-based acoustic likelihoods," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.293–296, April 1994.
[6] K.M. Berkling, T. Arai, and E. Bernard, "Analysis of phoneme based features for language identification," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.289–292, April 1994.
[7] T.J. Hazen and V.W. Zue, "Recent improvements in an approach to segment-based automatic language identification," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.1883–1886, April 1994.
[8] S. Kadambe and J.L. Hieronymus, "Language identification with phonological and lexical models," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.3507–3510, May 1995.
[9] Y. Yan, E. Barnard, and R.A. Cole, "Development of an approach to automatic language identification based on phone recognition," Comput. Speech Lang., vol.10, no.1, pp.37–54, Jan. 1996.
[10] J. Navratil and W. Zuhlke, "Phonetic-context mapping in language identification," Proc. European Conference on Speech Communication Technology, pp.71–74, Sept. 1997.
[11] Y. Yan and E. Barnard, "An approach to automatic language identification based on language-dependent phone recognition," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.3511–3514, May 1995.
[12] T. Arai, "Automatic language identification using sequential information of phonemes," IEICE Trans. Inf. & Syst., vol.E78-D, no.6, pp.705–711, June 1995.
[13] J. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," Proc. International Conference on Spoken Language Processing, pp.1283–1286, Oct. 2004.
[14] W. Shen, W. Campbell, T. Gleason, D. Reynolds, and E. Singer, "Experiments with lattice-based PPRLM language identification," Odyssey 2006 Speaker and Language Recognition Workshop, no.Li08R038, June 2006.
[15] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, R.A. Reynolds, and J.R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," Proc. International Conference on Spoken Language Processing, pp.89–92, Sept. 2002.
[16] W.M. Campbell, J.P. Campbell, D.A. Reynolds, and E. Singer, "Support vector machines for speaker and language recognition," Comput. Speech Lang., vol.20, no.2-3, pp.210–229, April 2006.
[17] L. Burqet, P. Metejka, and J. Cernocky, "Discriminative training techniques for acoustic language identification," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.209–212, May 2006.
[18] E. Noor and H. Aronowitz, "Efficient language identification using anchor models and support vector machines," Odyssey 2006 Speaker and Language Recognition Workshop, no.Li09R025, June 2006.
[19] H.Z. Li, B. Ma, and C.H. Lee, "A vector space modeling approach to spoken language identification," IEEE Trans. Speech Audio Process., vol.15, no.1, pp.271–284, Jan. 2006.
[20] C. White, I. Shafran, and J. Gauvain, "Discriminative classifiers for language recognition," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.213–216, May 2006.
[21] L.F. Zhai, M.H. Siu, X. Yang, and H. Gish, "Discriminatively trained language models using support vector machines for language identification," Odyssey 2006 Speaker and Language Recognition Workshop, no.Li06R022, June 2006.
[22] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," http://www.csie.ntu.edu.tw/˜cjlin/libsvm, 2001.
[23] T.F. Wu, C.J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," Journal of Machine Learning Research, vol.5, pp.975–1005, Aug. 2004.
[24] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika, "The OGI multilanguage telephone speech corpus," Proc. International Conference on Spoken Language Processing, pp.895–898, Oct. 1992.
[25] "HKUST Mandarin telephone speech," http://www.ldc.upenn.edu/Catalog/
[26] "The 2003 NIST LRE plan," http://www.nist.gov/speech/tests, 2003.
[27] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," Proc. DARPA Speech Recognition Workshop, pp.108–111, 1997.
[28] H. Gish, M.H. Siu, and R. Rohlicek, "Segregation of speaker for speech recognition and speaker identification," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.873–876, May 1991.
[29] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, pp.404–450, John Wiley & Sons, 1984.
[30] E. Barnard and Y. Yan, "Toward new language adaptation for language identification," Speech Commun., vol.21, no.4, pp.245–254, May 1997.
[31] N. Brummer and D. Leeuwen, "On calibration of language recognition scores," Odyssey 2006 Speaker and Language Recognition Workshop, no.Li14R049, June 2006.
[32] E. Singer, P. Torres-Carrasquillo, T. Gleason, W. Campbell, and D. Reynolds, "Acoustic, phonetic and discriminative approaches to automatic language recognition," Proc. European Conference on Speech Communication Technology, pp.1345–1348, 2003.
[33] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am., vol.87, no.4, pp.1738–1752, June 1990.
[34] A. Zolnay, R. Schlueter, and H. Ney, "Acoustic feature combination for robust speech recognition," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.457–460, March 2005.
[35] E. Barnard and R.A. Cole, "A neural-net training program based on conjugate-gradient optimization," Technical Report, no.CSE 89-014, Department of Computer Science, Oregon Graduate Institute of Science and Technology, 1989.



Hongbin Suo received his B.E. degree in Mechanical Engineering from Zhejiang University in 2002. Now he is a doctor student of ThinkIT Speech Lab, Institute of Acoustics (IOA), Chinese Academy of Sciences (CAS). His research is focused on automatic spoken language recognition.

Ming Li received his B.S. degree in electrical engineering from Nanjing University in 2001. He now is a master student of ThinkIT Speech Lab, IOA, CAS. His research interests include language identification, computational acoustics scene analysis, and audio watermarking

Ping Lu received her B.E. from University of electronic science and technology in 1996, and her Ph.D. in Electronic Engineering from Tsinghua University, 2003. She was working as a Postdoctoral Research Fellow in ThinkIT laboratory from 2003 to 2005. Now she is an associated professor in IOA, CAS, China. Her research field contains: Broadcast news corpus material identification, Speech information search etc. She has published over 20 papers in international or national learning periodicals and conferences. And, she has applied 1 national patent.

Yonghong Yan received his B.E. from Tsinghua University in 1990, and his Ph.D. from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998–2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Currently he is a professor and director of ThinkIT Lab. His research interests include speech processing and recognition, language/speaker recognition, and human computer interface. He has published more than 100 papers and holds 40 patents.