NIST 2007 Language Recognition Evaluation: From the Perspective of IIR*

Haizhou Li, Bin Ma, Kong-Aik Lee, Khe-Chai Sim, Hanwu Sun,
Rong Tong, Donglai Zhu, and Changhuai You

Institute for Infocomm Research,
Agency for Science, Technology and Research (A*STAR), Singapore
{hli,mabin,kalee,kcsim,hwsun,tongrong,dzhu,echyou}@i2r.a-star.edu.sg

Abstract. This paper describes the Institute for Infocomm Research (IIR) system for the 2007 Language Recognition Evaluation (LRE) conducted by the National Institute of Standards and Technology (NIST). The submitted system is a fusion of multiple state-of-the-art language classifiers using diversified discriminative language cues. We implemented several state-of-the-art algorithms using both phonotactic and acoustic features. We also investigated the system fusion and score calibration strategy to improve the performance of language recognition, and worked out a pseudo-key analysis approach to cross-validate the performance of the individual classifiers on the evaluation data. We achieve an equal-error-rate (EER) of 1.67 % on the close-set general language recognition test.

Keywords: Automatic spoken language recognition, NIST Language Recognition Evaluation, phonotactic features, acoustic features, fusion system, pseudo key.

1. Introduction

Automatic spoken language recognition (SLR) is a process of determining the identity of the language in a spoken document. As multilingual applications are demanded by the emerging need for globalization and the growing international business interflow, SLR has become an enabling technology in many applications such as multilingual conversational systems (Zue and Glass, 2000), multilingual speech recognition and translation (Waibel et al., 2000), and spoken document retrieval (Dai et al. 2003). It is also a topic of great importance in the areas of intelligence and security, where the language identities of recorded messages and archived materials need to be established before any information can be extracted. SLR technology also facilitates massive on-line language routing for voice surveillance over telephone network.

The National Institute of Standards and Technology (NIST) has conducted a series of evaluations of SLR technology in 1996, 2003, 2005 and 2007 (NIST, 2007). The language recognition evaluations (LREs) focus on language and dialect detection in the context of conversational telephony speech. They are conducted to foster research progress, with the goals of exploring promising new ideas in language recognition, developing advanced technology incorporating these ideas, and measuring the performance of this technology. The Institute for

22nd Pacific Asia Conference on Language, Information and Computation, pages 46–57

Infocomm Research (IIR) team has participated in the 2005 and 2007 NIST LREs and demonstrated the state-of-the-art technologies.

One of the fundamental issues in SLR is to explore the discriminative cues for spoken languages. In the state-of-the-art language recognition systems, these cues mainly come from the acoustic features (Sugiyama, 1991; Torres-Carassquilo et al., 2002; Burget et al., 2006; Campbell et al., 2006) and phonotactic representations (Hazen and Zue, 1994; Zissman, 1996; Berkling and Barnard, 1994; Corredor-Ardoy et al., 1997; Li and Ma, 2005; Ma, Li, and Tong, 2007), which reflect different aspects of spoken language characteristics. Another issue is how to effectively organize and exploit these language cues obtained from multiple sources in the recognition system design for the best performance.

Significant improvements in automatic speech recognition (ASR) have been achieved through exploiting the acoustic features representing the temporal properties of speech spectrum. These acoustic features, such as Mel-frequency Cepstral Coefficients (MFCCs), are also good choices to be the front-ends in language recognition systems. Gaussian mixture model (GMM), which can be seen as a one-state hidden Markov model (HMM) (Rabiner, 1989), is a simple modeling method to provide a multimodal density and is reasonably accurate when speech data are generated from a set of Gaussian distributions. It has demonstrated a great success in text-independent speaker recognition (Reynolds, Quatieri, and Dunn, 2000). In language recognition, GMM is also an effective method to model the unique characteristics among languages (Torres-Carassquilo et al., 2002). The support vector machine (SVM) has proven to be a powerful classifier in many pattern classification tasks. It is a discriminative classifier to separate two classes with a hyperplane in a high-dimensional space. The generalized linear discriminant sequence kernel (GLDS) has been proposed to apply SVM for speaker and language recognition (Campbell et al., 2006). The cepstral feature vectors extracted from an utterance are expanded to a high-dimensional space by calculating all the monomials.

In recent years, phonotactic features have been shown to provide effective cues for language recognition. The phonotactic features are extracted from an utterance to represent phonetic constraints in a language. Although common sounds are shared considerably across spoken languages, the statistics of these sounds, such as phone n-gram, can differ considerably from one language to another. Parallel Phone Recognizers followed by Language Models (PPR-LM) (Zissman, 1996) uses multiple parallel phone recognizers to convert the input utterance into a phone token sequence. It is followed by a set of *n*-gram phone language models that imposes constraints on phone decoding and provides language scores. Instead of *n*-gram phone language models, vector space modeling (VSM) was proposed as the classifier (Li, Ma, and Lee, 2007), called PPR-VSM. For each phone sequence generated from the multiple phone recognizers, the occurrences of phone *n*-grams are counted. A phone sequence is then represented as a high-dimensional vector of *n*-gram occurrence. SVM is used as the classifier on the concatenated n-gram occurrence vectors.

It is generally agreed upon that the integration with different cues of discriminative information can improve the performance of language recognition (Adda-Decker et al., 2003). The information extraction and organization of multiple sources has been critical to a successful language recognition system (Singer et al., 2003; Tong et al., 2006). In this paper, we will report our language recognition system submitted to the 2007 NIST LRE. The system is based on the fusion of multiple classifiers, each providing unique discriminative cue for language classification. In order to avoid a spoiled classifier in the submitted fusion system, we have designed a pseudo key analysis approach to check the integrity of each individual classifier before the system fusion.

The remainder of this paper is organized as follows. The evaluation data and evaluation metric of the 2007 NIST LRE will be introduced in Section 2. The system structure together with the phonotactic and acoustic language classifiers will be presented in Section 3. The fusion of multiple language classifiers and language recognition results on the 2007 NIST LRE evaluation data will be described in Section 4. The pseudo key analysis will be shown in Section 5. Finally in Section 6, we summarize our findings in language recognition.

2. Data and Metric

2.1. Evaluation Data

There are six test categories in the 2007 NIST LRE involving 26 target languages and dialects:

- General Language Recognition (LR) including 14 languages, Arabic, Bengali, Chinese, English, Hindustani, Spanish, Farsi, German, Japanese, Korean, Russian, Tamil, Thai and Vietnamese.
- Chinese LR including four Chinese dialects, Cantonese, Mandarin, Min and Wu.
- Mandarin Dialect Recognition (DR) including Mainland Mandarin and Taiwan Mandarin.
- English DR including American English and India English.
- Hindustani DR including Hindi and Urdu.
- Spanish DR including Caribbean Spanish and non-Caribbean Spanish.

Both closed-set and open-set tests in the six categories were conducted. For the closed-set tests, the non-target languages will be limited to those languages and dialects known to the system. For the open-set test the non-target languages will also include all other unknown languages such as Italian, Punjabi, Tagalog, Indonesian, and French. These unknown languages were not disclosed to participants, and the training data for these languages were not made available.

There are three test conditions to evaluate the system performance under different test segment durations:

- 3 seconds of speech (2-4 seconds actual)
- 10 seconds of speech (7-13 seconds actual)
- 30 seconds of speech (25-35 seconds actual)

The silence was not removed from speech so a segment could be much longer. There are 2510 segments for each of the three durations.

2.2. Training and Development Data

All the phonotactic and acoustic classifiers were trained with the LDC CallFriend corpus[1] and the LRE 2007 development databases released by NIST to all the participants. The phone recognizers used for phonotactic features were trained with OGI Multilingual database (Muthusamy, Cole, and Oshika, 1992) and IIR-LID database (Tong et al., 2006). The weights of fusion system were tuned on the LRE 1996, 2003, 2005 databases as well as the LRE 2007 development database.

2.3. Evaluation Metric

The primary evaluation metric is taken as the average cost performance C_{avg} (NIST LRE, 2007), which indicates the pair-wise language recognition performance, represented in terms of detection miss and false alarm probabilities, for all target/non-target language pairs. For the case of closed-set test condition, the C_{avg} is given by

$$C_{\text{avg}} = \frac{1}{N_{\text{tar}}} \sum_{l \in L_{\text{tar}}} \left\{ 0.5 P_{\text{miss}}(l) + 0.5 \times \frac{1}{(N_{\text{tar}} - 1)} \sum_{l' \in L_{\text{non}}} P_{\text{FA}}(l, l') \right\} \quad (1)$$

where L_{tar} is the set of N_{tar} target languages (e.g., $N_{\text{tar}} = 14$ for general LR). Notice that the miss probability P_{miss} is computed separately for each target language. All other languages are treated as non-target languages to compute the false alarm probabilities P_{FA} for each target/non-target language pairs. A complete definition of C_{avg} can be found in (NIST LRE, 200). In addition to the C_{avg}, we also report the results in terms of the average equal-error-rate (EER). That is, we compute the EER for each of the target language and take their average as the performance measure.

[1] http://www.ldc.upenn.edu/

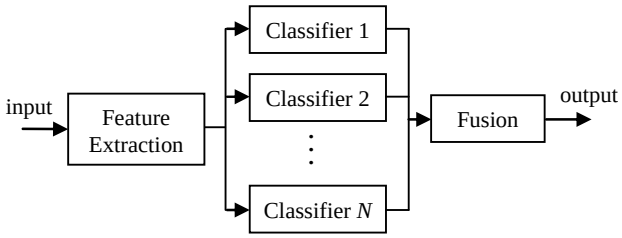

Figure 1: Fusion of multiple language classifiers.

3. System Description

The IIR system submitted to the 2007 NIST LRE is a fusion of multiple language classifiers. Figure 1 shows the overall framework.

3.1. Feature Extraction

The first stage of the feature extraction process is the Voice Activity Detection (VAD). Two types of VAD were used:

- Frame-based VAD

  For the acoustic classifiers, an energy based voice activity detector (VAD) is applied to remove silence frames and to retain only the high quality speech frames for language recognition. The frames whose energy level is more than 30dB below the maximum energy of the entire utterance are considered silence and therefore removed. Furthermore, if there are more than 40% of the frames are retained, only the top 40% of the frames with higher SNR are retained. The rest of the frames are discarded. There are approximately 30% of the frames which are actually selected for further processing.

- Segment-based VAD

  For phonotactic classifiers, segment-based VAD is used. Based on the VAD speech frame index obtained in the above, we first join continuous speech frames to form the speech segments. If the resulting segment is longer than 8 seconds, the segment is further split at the frame in that segment with the lowest energy. This is repeated until the resulting segment is less than 8 seconds in long. The final segments are padded with 200ms silence at both ends.

After VAD, two types of short time cepstral features, Mel Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs), are adopted as the basic features for acoustic classifiers. To capture temporal information across multiple frames, Shifted Delta Cepstral (SDC) coefficients (Torres-Carassquilo et al., 2002) are further applied to the frame-based MFCCs and LPCCs.

3.2. Phonotactic Classifiers

The phonotactic classifiers use multiple phone recognizers as the front-end to derive phonotactic statistics of a language. Since the individual phone recognizers are trained on different languages, they capture different acoustic characteristics from the speech data. Therefore, combining these recognizers together improves the overall language recognition performance.

The PPR front-end can be followed by both the phone *n*-gram language models (LM) (Zissman, 1996) and the vector space modeling (VSM) backend (Li, Ma, and Lee, 2007). The LM backend evaluates each token sequence using multiple language models, each of which describes a token sequence from the perspective of a target language. With VSM backend, the *n*-gram statistics from each token sequence form a high-dimensional feature vector, also known as a *bag-of-sounds* (BOS) vector (Li and Ma, 2005). A composite vector is constructed by stacking multiple *bag-of-sounds* vectors derived from multiple token sequences.

3.2.1. PPR-LM Classifier

With the PPR front-end, the backend of the language classifier can be language models for capturing the phonotactic constraints for each target language. PPR-LM approach (Zissman, 1996) uses the PPR front-end to convert a spoken utterance into multiple sequences of phones.

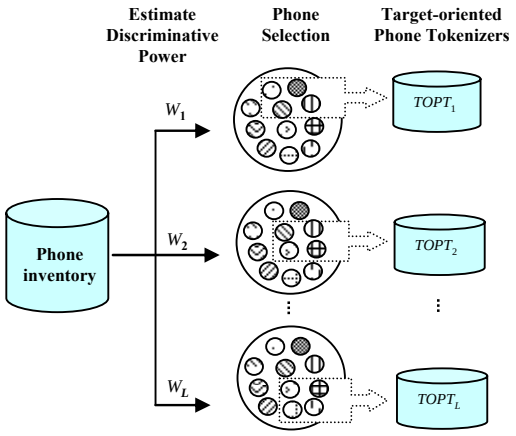

Figure 2: Construction of target oriented phone tokenizers.

Then a set of *L n*-gram phone language models estimates the likelihood phonotactic scores for the spoken documents in order to produce classification decisions.

3.2.2. PPR-VSM Classifier

Suppose that we have *F* phone recognizers with a phone inventory of $v=\{v_1,\ldots,v_\tau, \ldots,v_F\}$ and the number of phones in v_τ is n_τ. An utterance is decoded by these phone recognizers into *F* independent sequences of phone tokens. Each of these token sequences can be expressed by a high dimensional phonotactic feature vector with the *n*-gram counts. The dimension of the feature vector is equal to the total number of *n*-gram patterns needed to highlight the overall behavior of the utterance. If unigram and bigram are the only concerns, we will have a vector of $n_\tau+n_\tau^2$ phonotactic features, to represent the utterance by the τth phone recognizer.

For each target language, an SVM is trained by using the composite feature vectors in the target language as the positive set and the composite feature vectors in all other languages as the negative set. With *L* target languages, we project the high dimensional composite feature vectors into a discriminative feature vector with a much lower dimension (Ma, Li, and Tong, 2007).

We formulate the language recognition as a hypothesis test. For each target language, we build a language detector which consists of two GMMs $\{\lambda^+,\lambda^-\}$. The GMM trained on the discriminative vectors of the target language is called the positive model λ^+, while the GMM trained on those of its competing languages is called the negative model λ^-. We define the confidence of a test sample *O* belonging to a target language as the posterior odds in a hypothesis test under the Bayesian interpretation. We have H_0, which hypothesizes that *O* is language λ^+, and H_1, which hypothesizes otherwise. The posterior odd is approximated by the likelihood ratio $\Lambda(O)$ that is used for the final language recognition decision.

$$\Lambda(O)=\log\left(\frac{p(O|m^+)}{p(O|m^-)}\right) \tag{2}$$

3.2.3. Target-Oriented Phone Tokenizer (TOPT)

In the PPR framework, the languages of parallel phone recognizers, also known as phone tokenizers, and target languages may not have to be the same languages. For example, an English phone recognizer functions as a human listener of English background, trying to extract the discriminative information from the spoken utterances of each target language from its perspective. The discriminative information is expressed in an English phone sequence. In general, the performance gain increases with a greater number of parallel recognizers.

We proposed to design the target-oriented phone tokenizers (TOPTs) (Tong et al., 2008) rather to use the same phone recognizer for all the target languages in the PPR practice. For example, Arabic-oriented English phone tokenizer, Mandarin-oriented English phone tokenizer, as Arabic and Mandarin each is believed to have its unique phonotactic features to an English listener.

Note that not all the phones and their phonotactics in the target language may not provide equally discriminative information to the listener, it is desirable that the phones in each of the TOPTs can be those extracted from the full phone set of a phone recognizer, and having highest discriminative ability in distinguishing the target language from other languages.

The target-oriented phone selection strategy is illustrated in Figure 2. Assuming we have a language recognition task of L target languages, given a phone recognizer with phone inventory $v=\{v_1,v_2\cdots,v_i,\cdots,v_n\}$ which contains n phones, we estimate the discriminative power of each phone v_i in distinguishing a target language l_k from other target languages: l_j with $j\in[1,L]$ and $j\neq k$. The discriminative power of phones in v for distinguishing language l_k from others can be denoted as $W_k=\{w_{v_1,k}w_{v_2,k}\cdots w_{v_m,k}\}$. We select a subset of phones that have highest discriminative power to construct a new target-oriented phone tokenizer, $TOPT_k$. In this way, we can construct L new target-oriented phone tokenizers, one for each target language.

3.2.4. Phonetic and Acoustic Diversifications (PAD)

Phonetic and acoustic diversifications may be applied to both PPR-LM and PPR-VSM systems. The conventional approach adopts phonetic diversification, where the parallel phone recognizers are trained on speech data from different languages with different phone sets. On the other hand, we proposed an alternative methodology where phone recognizers using different acoustic models trained on the same speech data with the same phone set (Sim and Li, 2007, 2008) are used to achieve acoustic diversification. Analogous to system combination for speech recognition in which merging outputs from multiple systems with different error patterns helps to improve the final performance, using multiple acoustic models aims to form the contractive parallel phone recognition systems using different modeling techniques and training paradigms, without requiring additional phonetically transcribed speech data.

3.3. Acoustic Classifiers

Acoustic classifiers exploit acoustic features directly. There are two main approaches, Gaussian mixture modeling (GMM) on short-time cepstral features, such as MFCCs, LPCCs, and the Shifted Delta Cepstral (SDC) coefficients, and support vector machine (SVM) modeling on high dimension acoustic features, such as the polynomial expansion of short-time cepstral features.

3.3.1. MMI-GMM

In the standard Maximum Likelihood (ML) training framework for GMM, the objective function is to maximize the total log likelihood of training data:

$$\mathrm{F}_{\mathrm{ML}}(\theta)=\sum_{r=1}^{R}\log p(O_r|s_r) \tag{3}$$

where θ is the model parameter set and O_r is the rth observation sequence, R denotes the total number of training utterances, and s_r is the correct language identity of the rth utterance. The ML estimation maximizes the likelihood of each model generating the training data independently.

The discriminative training techniques have been successfully applied in large vocabulary continuous speech recognition (LVCSR) systems. One of the most popular discriminative training approaches, *maximum mutual information* (MMI) training, has been proved to efficient in the Gaussian mixture modeling for language recognition (Bueget, Matejka, and Cernocky, 2006). The objective function of MMI is posterior probability of correctly recognizing all training utterances. It estimates the GMM parameters in a discriminative manner by maximizing the following objective function:

$$\mathrm{F}_{\mathrm{MMI}}(\theta)=\sum_{r=1}^{R}\log\left(\frac{p_\theta(O_r|s_r)P(s_r)}{\sum_{\forall s}p_\theta(O_r|s)P(s)}\right) \tag{4}$$

where $P(s_r)$ and $P(s)$ are the prior terms and we consider the prior probabilities of all languages equal. The denominator $\sum_{\forall s}p_\theta(O_r|s)P(s)$ is the likelihood of utterance O_r given the competing language models.

3.3.2. GLDS Kernel

SVM has been proven to be an effective two-class classifier for pattern classification problems. To adopt SVM for classification of speech utterances is not straightforward since speech utterances are often parameterized as variable-length sequences of cepstral feature vectors. A kernel function that can measure the similarity between two sequences of speech feature vectors has to be constructed. The *generalized linear discriminant sequence* (GLDS) kernel has been proposed for speaker and language recognition (Campbell et al., 2006) on acoustic feature vectors. Given two sequences, $X=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and $Y=\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, of feature vectors, the GLDS kernel is given by

$$K_{\text{GLDS}}(X,Y)=\mathbf{b}_x^T\mathbf{R}^{-1}\mathbf{b}_y \tag{5}$$

where m and n denote the number of feature vectors in the sequences X and Y, respectively. In (5), the two sequences become comparable by mapping them to a high-dimensional vector space via

$$\mathbf{b}_x=\frac{1}{m}\sum_{\mathbf{x}\in X}\tilde{\mathbf{b}}(\mathbf{x}) \text{ and } \mathbf{b}_x=\frac{1}{n}\sum_{\mathbf{y}\in Y}\tilde{\mathbf{b}}(\mathbf{y}) \tag{6}$$

where $\tilde{\mathbf{b}}(\cdot)$ denotes the polynomial expansion function. For $\mathbf{x}=[x_1, x_2]^T$ and considering all monomials up to the second order, the expansion function is given by $\tilde{\mathbf{b}}(\mathbf{x})=[1, x_1, x_2, x_1^2, x_1x_2, x_2^2]^T$. In our final implementation, we used all monomials up to the third order. In (5), $\mathbf{R}=(\mathbf{U}^T\mathbf{U})/N_U$ is a correlation matrix calculated from a data matrix $\mathbf{U}$ that consists of the expansions of the entire set of N_U training feature vectors. For computational simplicity, it is customary to assume that the matrix $\mathbf{R}$ is diagonal. An SVM is then constructed as the sum of kernel functions in the following form

$$f(X)=\sum_l \alpha_l K_{\text{GLDS}}(X_l, X)+\beta \tag{7}$$

Here, $\{X_l\}$ denotes the support vectors, β is the bias, and the term α_l, for $\sum_l \alpha_l=0$, $\alpha_l>0$, indicates the weight of the lth support vector in the expanded feature space.

3.3.3. Probabilistic Sequence Kernel (PSK)

The PPR (see Section 3.2) serves as a front-end decoder that extracts phonotactic information (i.e., phone sequences) from which the speech utterance can be characterized in terms of the occurrence and co-occurrence statistics of various phones. In (Lee, You, and Li, 2008), we explored the use of acoustically-defined units, instead of the linguistically-defined phones, in characterizing speech utterances and spoken languages. In particular, we train an ensemble of acoustic sound classes in a self-organized manner, each modeled with a Gaussian distribution, to form a speech sound inventory analogous to the phone inventory. We interpret the acoustic sound classes to represent some general vocal tract configurations in producing various speech sounds. The self-organized nature of these acoustic sound classes circumvents the need of laborious phonetic transcription. Furthermore, the structural simplicity of the Gaussian distributions allows us to train sufficient number of acoustic units to transcribe the sound of spoken languages in an effective manner.

We formulate the acoustic sound inventory in a form of sequence kernel, referred to as the *probabilistic sequence kernel* (PSK), for SVM. Similar to that of the GLDS kernel mentioned earlier, the PSK maps variable-length utterances into fixed- and high-dimensional vectors in order to transform a complex classification task into a linearly separable one in a higher-dimensional vector space. Let $p(\mathbf{x}|j)\sim \mathrm{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, for $j=1,2,\ldots,L$, denote the inventory of acoustic sound classes. Using these sound classes as bases, the feature expansion is defined as

$$\tilde{\mathbf{p}}(\mathbf{x})=[p(j=1|\mathbf{x}),\ p(j=2|\mathbf{x}),\ \ldots,\ p(j=L|\mathbf{x})]^T \tag{8}$$

where $p(j|\mathbf{x})$ denotes the posterior probability of the jth acoustic class (the prior probability of each acoustic class is determined during the training stage as noted below). Each element of the expansion $\tilde{\mathbf{p}}(\mathbf{x})$ gives the probability of occurrence of the jth acoustic class evaluated for a given feature vector $\mathbf{x}$. The average probabilistic count across the entire sequence X is given by

$$\mathbf{p}_x = \frac{1}{m}\sum_{\mathbf{x}\in X}\tilde{\mathbf{p}}(\mathbf{x}). \tag{9}$$

The vector $\mathbf{p}_x$ can be interpreted as an M-bin histogram indicating the probabilities of occurrence of various acoustic sound classes observed in the given speech utterance X. Given two sequences, the PSK measures their similarity as the inner product between their expanded vectors, $\mathbf{p}_l$ and $\mathbf{p}_x$, as follows

$$K_{\text{PSK}}(X,Y) = \mathbf{p}_x^T \mathbf{R}^{-1} \mathbf{p}_y. \tag{10}$$

Compared to the GLDS kernel (5), the PSK hinges on the prior knowledge that the frequency of occurrence of speech sounds differs from one language to another in establishing the bases. This prior knowledge is not exploited in the GLDS kernel, leading to some performance deficiency.

4. Fusion of Classifiers

This section describes the fusion strategy for the IIR submission to the NIST 2007 Language Recognition Evaluation (LRE07). The final submitted system is a linear fusion of the scores contributed by ten individual classifiers. These classifiers are summarized in Table 1.

Half of the classifiers are phonotactic classifiers while the remaining halves are acoustic classifiers. Two novel PPR-VSM classifiers were introduced to the LRE07 submission, namely the TOPT and PAD classifiers (see Sections 3.2.3 and 3.2.4 respectively). In addition, our system also made use of the HMM/NN hybrid phone recognizers provided by the Brno University of Technology (BUT)[2]. On the other hand, PSK, a novel acoustic classifier with generative front-end was also used (see Section 3.3.3). Furthermore, two GLDS acoustic classifiers were built using the MFCC and LPCC features. Two GMM classifiers were also trained using the ML and MMI criteria.

The final system was obtained by means of linear fusion of the scores from the ten individual classifiers:

$$s_i = \sum_{c=1}^{C} w_c s(c,i) + b \tag{11}$$

where C is the total number of classifiers and $s(c,i)$ is the score of the ith trial from the cth classifier. The fusion parameters consist of the classifier specific weights w_c and the global bias b. Two objectives were used to tune the fusion parameters:

a. minEER:

$$(w_c, b)_{\text{minEER}} = \min_{w_c, b} \left\| P_{miss} - P_{FA} \right\| \tag{12}$$

where the miss and false alarm probabilities are given by

$$\begin{aligned} P_{\text{miss}} &= \frac{\left|\{i : i \in \text{True}, s_i < b\}\right|}{\left|\{i : i \in \text{True}\}\right|} \\ P_{\text{FA}} &= \frac{\left|\{i : i \in \text{False}, s_i \geq b\}\right|}{\left|\{i : i \in \text{False}\}\right|} \end{aligned} \tag{13}$$

and $|\{\ldots\}|$ denotes the cardinality of the set.

b. Logistic Linear Regression (LLR):

$$(w_c, b)_{\text{LLR}} = \max_{w_c, b} \sum_{\forall i} \left(\frac{1}{1+\exp(-y_i s_i)} \right) \tag{14}$$

where

[2] http://www.fit.vutbr.cz/research/groups/speech/index_e.php?id=phnrec

Table 1: List of 10 individual classifiers used in the IIR NIST 2007 Language Recognition Evaluation submission.

Phonotactic Classifiers	Acoustic Classifiers
PPR-VSM	PSK
TOPT-PPR-VSM	MFCC-GLDS
PAD-PPR-VSM	LPCC-GLDS
BUT-PPR-LM	ML-GMM
BUT-PPR-VSM	MMI-GMM

Table 2: $\mathbf{C_{avg}}$ performance using the minEER+LLR fusion method for the General LR closed-test tasks.

Systems	C_{avg} (%)		
	30s	10s	3s
Worst individual	10.23	18.16	33.05
Best individual	3.54	9.22	20.59
Fusion	2.75	6.15	16.40

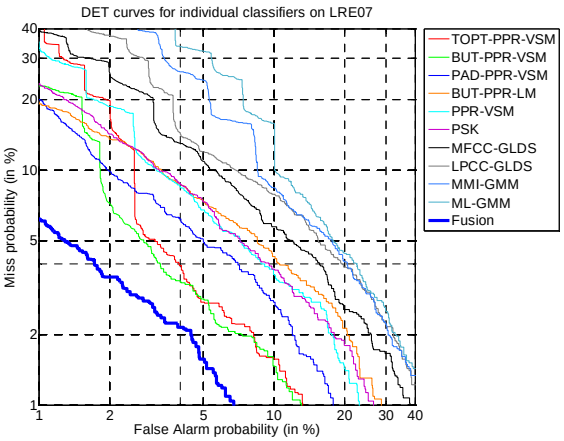

Figure 3: DET curves of individual classifiers and the final fusion system for the 30s General LR closed-test task.

$$y_i = \begin{cases} 1, & i \in \text{True} \\ 0, & i \in \text{False} \end{cases} \tag{15}$$

LLR attempts to transform the scores from multiple classifiers to the log likelihood ratios. The LLR is performed using the FoCal toolkit[3].

The final fusion parameters were obtained as the average of the parameters estimated using the above objectives, i.e.,

$$\begin{aligned} w_c &= \frac{1}{2}[(w_c)_{\text{minEER}} + (w_c)_{\text{LLR}}] \\ b &= \frac{1}{2}[(b)_{\text{minEER}} + (b)_{\text{LLR}}] \end{aligned} \tag{16}$$

The fusion parameters were calibrated on the development data comprising the NIST 1996, 2003 and 2005 evaluation sets as well as the 2005 OHSU development data.

4.1.1. Fusion results

Figure 3 shows the Detection Error Trade-off (DET) curves for the 10 individual classifiers as well as the final fusion system for the 30s General LR closed-test task. The top 3 performing classifiers include the BUT-PPR-VSM, TOPT-PPR-VSM and PAD-PPR-VSM classifiers.

The C_{avg} performance of the best and worst individual classifiers as well as the fusion system for the 30s, 10s and 3s General LR closed-test tasks is summarized in Table 2. The relative improvements obtained from fusion over the best individual classifier were 22.3%, 33.3% and 20.3% for the 30s, 10s and 3s tasks respectively.

4.1.2. Open-test versus Closed-test

Table 3 shows the comparison of the *EER (%)* and C_{avg} (%) performance for the open-test and closed-test conditions on various tasks. In general, it was found that the General LR tasks are relatively easier compared to the Chinese LR and the other dialect recognition (DR) tasks. In particular, the Hindustani DR and Spanish DR tasks were the hardest, with C_{avg} performance greater than 30%. As expected, the performance of the closed-test tasks is generally better than that of the open-test tasks due to the presence of the out-of-set languages in the open-test

[3] http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm

condition. Note that the C_{avg} performance depends on the decision threshold which may not coincide with the EER operating point. There are several cases (e.g. Hindustani DR and Spanish DR) where the C_{avg} performance for the open-test condition outperformed the closed-test condition due to the poor decision threshold in the later condition. The decision thresholds for the open-test conditions were estimated using development data that contains some out-of-language (OOL) languages to learn the appropriate trade-off between false acceptance (false alarm) and false rejection (miss). This has been found to yield improved performance compared to using data without OOL languages. For example, the C_{avg} performance for the 30s General LR open-test task would have been 5.71% instead of 4.28% if the decision threshold was tuned using development data without OOL languages.

Table 3: Comparison of EER and C_{avg} performance for the open-test and closed-test conditions on various tasks

Systems	Test Conditions	30s		10s		3s	
		EER	C_{avg}	EER	C_{avg}	EER	C_{avg}
General LR	Closed-test	1.67	2.75	5.87	6.15	15.38	16.40
	Open-test	2.34	4.28	6.79	8.20	15.92	17.88
Chinese LR	Closed-test	4.90	5.99	8.30	9.51	19.01	20.96
	Open-test	4.89	5.96	9.03	8.02	21.71	18.29
Mandarin DR	Closed-test	12.66	12.72	24.69	24.45	29.74	31.70
	Open-test	15.83	13.39	24.05	19.89	36.07	30.03
English DR	Closed-test	9.38	17.34	14.38	23.13	23.75	24.06
	Open-test	11.25	14.59	16.88	18.58	26.88	26.45
Hindustani DR	Closed-test	32.34	31.56	35.00	34.84	41.09	41.72
	Open-test	35.16	29.12	39.06	32.40	43.75	38.15
Spanish DR	Closed-test	27.97	34.38	33.12	40.00	42.50	44.06
	Open-test	32.66	30.28	40.00	33.50	42.50	37.88

5. Pseudo Key Analysis

We apply a pseudo-key analysis scheme to cross validate the performances of individual classifiers. It is to find out the abnormal classifier and prevent the error in the final fusion system without knowing the true keys of evaluation data. Suppose that the ratio of genuine/imposter test trials is around *1*:(*L-1*), where *L* is the number of the target languages. From the pool of scores of *M* trials from each classifier *c*, we choose *M*/*L* trials with the highest scores as genuine trials and the remaining trials as impostor trials, i.e.,

$$\tilde{k}(c,i)=\begin{cases}\text{True}, & \text{if } s(c,i)\geq\tilde{T}_c\\ \text{False}, & \text{if } s(c,i)<\tilde{T}_c\end{cases} \tag{17}$$

where $\tilde{k}(c,i)$ denotes the pseudo key for the *i*th trial of the *c*th classifier and the threshold $\tilde{T}_c$ is set such that there are *M/L* trials whose scores are above it. In the above equation, $s(c,i)$ represents the score of the *i*th trial from the *c*th classifiers. Using the pseudo keys from all classifiers, we compute the pseudo EER for the *c*th classifier as

$$EER_{\text{Pseudo}}(c)=\frac{1}{N-1}\sum_{g=1,g\neq c}^{N} EER(c|g) \tag{18}$$

where

$$EER(c|g)=\alpha(s(c,i)|\tilde{k}(g,i),i=1,M) \tag{19}$$

is the operator computing the EER of *c*th classifier using the psuedo keys obtained from the *g*th classifier, and *N* is the total number of classifiers.

We found that the genuine and imposter scores can be roughly expressed as two Gaussian distributions. The probability of error with the pseudo keys obtained from the *c*th classifier is given by

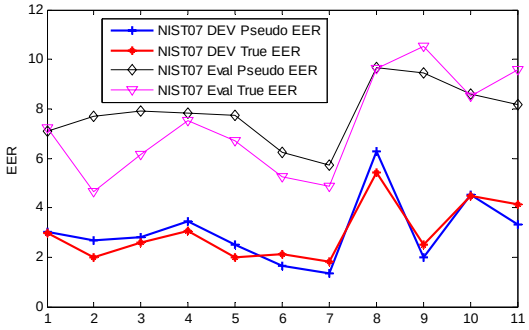

Figure 4: Pseudo and actual EERs evaluated on the development and evaluation sets of the NIST LRE 2007 (30s General LR close-test condition).

$$P(error,c)=P\left(s<\tilde{T}_c|m_1,\sigma_1\right)+P\left(s\geq\tilde{T}_c|m_0,\sigma_0\right) \tag{20}$$

where $\{m_1,\sigma_1\}$ and $\{m_0,\sigma_0\}$ are the mean and variance of the genuine and imposter score distributions, $\mathrm{N}(s;m_1,\sigma_1)$ and $\mathrm{N}(s;m_0,\sigma_0)$, and $\tilde{T}_c$ is the threshold defined in (17). Obviously, the error probability is the overlapped sections of the two distributions as indicated in (20). The performance of each classifier depends on the area of this overlapped section. The smaller the overlapped section, the better the classifier is. When this overlapped section is minimized, the classifier achieves desired performance. An outlier classifier will give a large overlap between the genuine and imposter distributions, resulting in high error rate with respect to pseudo keys.

We used the pseudo-key approach to analyze the performance of individual classifierson the LRE07 development and evaluation data sets. The pseudo EERs were computed using (17) and (18). Figure 4 compares the pseudo and actual EERs for all the classifiers. It is shoen that there exists a consistency between the pseudo and actual EERs on both the development and evaluation sets. The pseudo EERs can therefore provide a rough performance indication of the classifiers.

6. Discussion

A description of a language recognition system has been presented as it was developed for the 2007 NIST LRE. The submission was built upon multiple classifiers using generative and discriminative classification techniques, and was purposely designed to exploit the benefits of both phonotactic and acoustic features. Notably, we introduced three novel language classifiers, two phonotactic and one acoustic, in our LRE07 submission. The TOPT and PAD classifiers were shown to be successful refinements to the conventional phonotactic approach. On the other hand, the PSK bridges the gap between acoustic and token-based techniques. All the classifiers were combined at the score level with a simple linear fusion giving an EER of 1.67 % and a C_{avg} of 2.75 % under the general LR core test condition. The LRE results represent the state-of-the-art performance with an effective design and implementation.

7. References

Zue, V. W. and J. R. Glass, “Conversational interfaces: advances and challenges,” Proc. IEEE, vol. 88, no. 8, pp. 1166-1180, 2000.

Waibel, A., P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, “Multilinguality in speech and spoken language systems,” *Proc. IEEE*, vol. 88, no. 8, pp. 1181-1190, 2000.

Dai P., U. Iurgel, and G. Rigoll, “A novel feature combination approach for spoken document classification with support vector machines,” in *Proc. Multimedia Information Retrieval Workshop*, 2003.

National Institute of Standards and Technology. http://www.nist.gov/speech/tests/lang/2007/.

Sugiyama, M., “Automatic language recognition using acoustic features,” in *Proc. ICASSP*, 1991.

Torres-Carassquilo, P. A., E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in *Proc. ICSLP*, 2002.

Burget, L., P. Matejka, and J. Cernocky, “Discriminative training techniques for acoustic language identification,” in *Proc. ICASSP*, 2006, pp. I-209-212

Campbell, W. M., J. P. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo “Support vector machines for speaker and language recognition,” *Computer Speech and Language*, vol. 20, pp. 210-229, 2006.

Hazen, T. J. and V. W. Zue, “Recent improvements in an approach to segment-based automatic language identification,” in *Proc. ICASSP*, 1994.

Zissman, M. A., “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44, 1996.

Berkling, K. M. and E. Barnard, “Analysis of phoneme-based features for language identification,” in *Proc. ICASSP*, pp. 289-292, 1994.

Corredor-Ardoy, C., J. L. Gauvain, M. Adda-Decker, and L. Lamel, “Language identification with language-independent acoustic models,” in *Proc. Eurospeech*, 1997.

Li, H. and B. Ma, “A phonotactic language model for spoken language identification,” in *Proc. ACL*, 2005.

Ma, B., H. Li, and R. Tong, "Spoken Language Recognition Using Ensemble Classifiers", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2053-2062, Sep. 2007.

Rabiner, L. R., “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol.77, no.2, pp. 257-286, 1989.

Reynolds, D. A., T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Modeling,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

H. Li, B. Ma, and C.-H. Lee, “A vector space modeling approach to spoken language identification,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 271-284, 2007.

Adda-Decker, M., et al., “Phonetic knowledge, phonotactics and perceptual validation for automatic language identification,” in *Proc. ICPhS*, 2003.

Singer, E., P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, “Acoustic, phonetic and discriminative approaches to automatic language recognition,” in *Proc. Eurospeech*, 2003.

Tong, R., B. Ma, D. Zhu, H. Li, and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *Proc. ICASSP*, 2006.

Muthusamy, Y. K., R. A. Cole, and B. T. Oshika, “The OGI multi-language telephone speech corpus,” in *Proc. ICSLP*, 1992.

The 2007 NIST Language Recognition Evaluation plan, http://www.nist.gov/speech/tests/lang/2007/LRE07EvalPlan-v8b.pdf.

Tong, R., B. Ma, H. Li, and E. S. Chng, "Target-oriented phone tokenizers for spoken language recognition," in *Proc. ICASSP*, 2008, pp. 4221-4224.

Sim, K. C. and H. Li, “Fusion of contrastive acoustic models for parallel phonotactic spoken language identification”, in *Proc. Interspeech*, 2007, pp. 170-173.

Sim, K. C. and H. Li, “On acoustic diversification front-end for spoken language recognition”, to appear in *IEEE Trans. Audio, Speech and Language Processing*.

Bueget, L., P. Matejka, and J. Cernocky, “Discriminative training techniques for acoustic language identification”, in *Proc. ICASSP*, 2006, pp. 209-212.

Lee, K. A., C. You, and H. Li, “Spoken language recognition using support vector machines with generative from-end,” in *Proc*. ICASSP, 2008, pp. 4153-4156.