

Language Recognition Based on Score Distribution Feature Vectors and Discriminative Classifier Fusion

Jinyu Li¹, Sibel Yaman¹, Chin-Hui Lee¹, Bin Ma², Rong Tong², Donglai Zhu² and Haizhou Li²

¹School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA. 30332 USA
{jinyuli, syaman, chl}@ece.gatech.edu

²Institute for Infocomm Research, Singapore
{mabin, tongrong, dzhu, hli}@i2r.a-star.edu.sg

Abstract

We present the GT-IIR language recognition system submitted to the 2005 NIST Language Recognition Evaluation. Different from conventional frame-based feature extraction, our system adopts a collection of broad output scores from different language recognition systems to form utterance-level score distribution feature vectors over all competing languages, and build vector-based spoken language recognizers by fusing two distinct verifiers, one based on a simple linear discriminant function (LDF) and the other on a complex artificial neural network (ANN), to make final language recognition decisions. The diverse error patterns exhibited in individual LDF and ANN systems facilitate smaller overall verification errors in the combined system than those obtained in separate systems.

1. Introduction

NIST (National Institute of Standards and Technology) has coordinated evaluations of automatic language recognition technologies in 1996, 2003 and, recently, in 2005 to promote spoken language recognition research. Several techniques have achieved recent successes. The most popular framework is parallel phone recognition followed by language model (P-PRLM) [1]. It uses multiple sets of phone models to decode spoken utterances into phone sequences, and builds one set of phone language model (LM) for each P-PRLM tokenizer-target language pair. The P-PRLM scores are computed from language scores and the language with the maximum combination score is determined to be the recognized language. Another recently proposed approach is to use bag-of-sounds (BOS) models of phone-like units, such as acoustic segment units [2], to convert utterances into text-like documents. Then vector-based techniques, such as Gaussian mixture model (GMM) and support vector machine (SVM), can easily be adopted for language recognition [2] [3] [4].

The GT-IIR language recognition system submitted to the 2005 NIST language recognition evaluation (LRE) grows out of a collaborative effort between Georgia Institute of Technology (GT) and Institute for Infocomm Research (IIR). The system takes advantage of recent advances in P-PRLM and BOS frameworks and uses their corresponding models to obtain scores for all target languages, and concatenate them to form utterance-level score vector as front end feature, and train vector-based classifiers to perform spoken language recognition. Such feature vectors represent score distribution over all competing classes, and have been demonstrated to be effective in isolated word recognition, especially when

discriminative training techniques are incorporated into building the corresponding classifiers [5] [6]. Two distinct vector-based classifiers are considered. One is based on linear discriminant function (LDF) [7] and the other on artificial neural network (ANN) [8].

The LDF classifiers used in our system are obtained based on minimum classification error (MCE) [9] training, which has achieved a great success in automatic speech recognition. Our motivation for using MCE is to enhance the separation between LDF models of a target language and their competing languages. This is critical when there are only relatively few parameters used in a classifier, such as LDF. On the other hand the ANN classifiers are more complex in structure than the LDF ones, and can potentially cause over-fitting problems when there are not enough training samples.

Keeping in mind the advantages and shortcomings of each classifier, we are motivated to investigate new directions in fusing confidence scores generated by multiple classifiers to make final language recognition decisions. In our system, the combination verifier is formed with a simple linear weighting of confidence scores of LDF and ANN classifiers. The diverse error patterns exhibited in individual LDF and ANN systems result in smaller overall verification errors in the combined system than those obtained in separate systems.

2. NIST Language Recognition Evaluation 2005

GT and IIR co-operated together and contributed the GT-IIR system in NIST language recognition evaluation 2005. According to the evaluation specification [10], the system to be evaluated must determine whether or not the speech is from the target language/dialect given a test segment of speech and a target language/dialect. The target languages and dialects include American English, Indian-accented English, Hindi, Japanese, Korean, Mainland Mandarin, Taiwanese Mandarin, Spanish and Tamil.

The speech segments contain three nominal durations of speech, namely 3 seconds, 10 seconds and 30 seconds. The performance of a detection system is measured by a detection cost function C_{det} formulated for the i^{th} language as

$$C_{Det}(i) = C_{Miss} P_{Miss(i)|T arg et} P_{T arg et} + \frac{1}{N-1} \sum_{j \neq i} C_{FalseAlarm} P_{FalseAlarm(i)|NonT arg et(j)} (1 - P_{T arg et})$$

where N is the number of languages and $P_{\text{Target}}=P_{\text{NonTarget}}=0.5$. The final evaluation score is the average over all target languages.

3. System Description

We submitted three sets of results to the 2005 NIST LRE. The first system is for primary language recognition, and the second and third are for dialect detection. We obtained models for 16 languages/dialects in the training stage. They are Arabic, Farsi, French, German, Hindi, Japanese, Korean, Tamil, Vietnamese, 3 English dialects, 2 Mandarin dialects and 2 Spanish dialects. A 16-language/dialect training database consists of the Indian English dialect data from the IIR-LID [11] database and the other 15 languages/dialects from the CallFriend corpus [12].

A block diagram of the overall system is shown in Figure 1. We used two sets of score distribution features to represent each utterance. First a set of 112 scores are computed with the P-PRLM method, and then another set of 120 scores are evaluated with the BOS approach. These scores are concatenated to form a 232-dimension feature vector that is fed into the back end system, which is composed of several components. In the first stage, a confidence score is computed for each language in the ANN and LDF component verifiers. These scores are fused in the *Fusion Module*. Finally, a TRUE/FALSE verification decision is made and an individual confidence score is generated for each target language in the *Decision Module*. We now describe each module in detail in the following subsections.

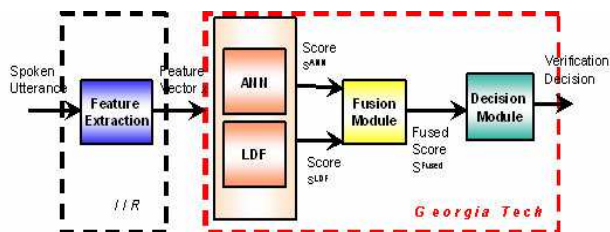


Figure 1: System block diagram for the GT-IIR system.

3.1. Score Distribution Feature Vector

Instead of using frame-based vectors as the front end features in most conventional LID systems, we extract utterance-based score vectors generated by P-PRLM and BOS models.

Seven phone recognizers were built: English, Korean, Mandarin, Japanese, Hindi, Spanish and German. English phonemes are trained from IIR-LID [11] corpus. Korean phonemes are trained from LDC Korean corpus (LDC2003S03). Mandarin phonemes are trained from the MAT corpus [13]. Other phonemes are trained from OGI-TS corpus [14]. 39-dimensional MFCC features are extracted from each frame. Utterance based cepstral mean subtraction is applied to the MFCC features to remove channel distortion. Each phoneme in the languages are modeled with a HMM of 3-state.

First, the 16-language/dialect training database is tokenized into a collection of text-like phone sequences from each of the 7 tokenizers. We compute P-PRLM scores based on the resulting phone sequences. This way, we train up to 3-gram phone LM for each P-PRLM tokenizer-target language pair, resulting in $16*7=112$ LMs. For each input utterance, 112

interpolated scores were derived to form a vector. In this way, all training utterances can be represented by a collection of 112-dimension score vectors.

Next, BOS scores were evaluated. The BOS method uses a universal sound recognizer to tokenize an utterance into a phone sequence, which is then converted into a count vector, known as BOS vector [2]. The universal sound inventory is a combined phoneme set from 6 languages: English, Mandarin, Japanese, Hindi, Spanish and German, a subset of the 7 languages above. There are 258 phonemes in total. For each phone sequence generated from the universal sound tokenizer, we count the occurrence of bi-phones. A phone sequence is then represented as a vector of bi-phone occurrence with $66,564 = 258 \times 258$ elements. A SVM is used to partition the high dimensional vector space. As SVM is a 2-way classifier, we train pair-wise SVM classifiers for the 16 target languages, resulting in $16*15/2=120$ SVM classifiers. The linear kernel is adopted when using SVM-light tool.

Finally, the above two sets of scores for each utterance, a vector of 112 dimensions obtained from the P-PRLM and a vector of 120 dimensions from the BOS methods, were concatenated to form the feature vector of 232 dimensions, x , for our back end system.

3.2. ANN Verifier

We trained a single feed-forward multilayer perceptron using a back-propagation procedure [8]. The network was structured to have 232 input layer nodes and 100 hidden layer nodes. The output layer has 16 nodes, one for each of the 16 target languages/dialects described above. A soft-max function was imposed at the output of the neural network so that output values $g_j(x)$, $j=1\dots 16$, simulate the *a posteriori* probability of each language/dialect given the input vectors. Hence the j th output value gives us a confidence measure about the language/dialect j characterizing the input x . In case of multiple outputs for a language (with multiple dialects), the dialect with the maximum output values was the selected language.

3.3. MCE Optimized LDF Verifier

Similar to the ANN verifier, 16 sets of LDFs were trained, one for each of the 16 languages and dialects. We performed discriminative training to minimize the average of the false-alarm rate FA_i and false rejection rate FR_i , $i=1,\dots,16$. A smooth approximation of the empirical error counts was imposed and a generalized probabilistic descent (GPD) [9] algorithm was used to update the classifier weights w in the $(k+1)$ st iteration as

$$w^{k+1} = w^k - \epsilon^k \left[\sum_{i=1}^{16} \nabla_w FA_i + \sum_{i=1}^{16} \nabla_w FR_i \right]$$

The first step for this approximation is picking a linear discriminant function $g_i(x, w)$ and an anti-discriminant function $G_i(x, w)$ for each language

$$g_i(x, w) = w_i^T x + w_{i0}, i = 1, \dots, 16$$

$$G_i(x, w) = \log \left[\frac{1}{15} \sum_{i \neq j} \exp(\eta g_j(x, w)) \right]^{1/\eta}$$

The misclassification measure $d_i(x, w) = -g_i(x, w) + G_i(x, w)$ leads to approximating the class loss function as

$$l_i(x, w) = \frac{1}{1 + \exp(-\alpha_i d_i(x, w) + \beta_i)}$$

The false-alarm rates FA_i and false-rejection rates FR_i for the i th language can now be expressed as

$$FA_i = \frac{1}{|\Omega - \Omega_i|} \sum_{x \in \Omega_i} (1 - l_i(x, w)) \quad FR_i = \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} l_i(x, w)$$

3.4. Language Confidence Scores

Two verifiers, one based on ANN and the other on LDF, were used to compute a pair of confidence scores for each of these 16 languages and dialects. Note that only 7 out of the 12 primary languages were of interest in LRE05. The remaining 5 languages were used to model the class ‘others’ given that the systems can potentially be challenged by languages other than the listed ones. In fact, in the LRE05 test set, only German was included as ‘others’. In the meantime, only 2 out of 3 English dialects, and both Mandarin dialects were of interest.

If we were to design a language identification system, picking the largest score among the competing ones would be the best procedure to follow. However, since a verification system requires a TRUE/FALSE decision for each target language, we need to calculate a *relative* confidence score, $s_j(x)$, representing the odds of each language. Hence, for every target language j , we compute the *language confidence score*:

$$s_j(x) = g_j(x) - \log \left[\frac{1}{N-1} \sum_{i \neq j} \exp(g_i(x)\eta) \right]^{1/\eta} \quad j = 1 \dots 7, \quad i = 1 \dots 8 \quad (1)$$

For dialect identification, the language confidence score was calculated as the difference between the target and non-target dialect outputs, i.e.

$$s_i(x) = g_i(x) - g_j(x), \quad i, j = 1, 2.$$

After we computed the two scores for each language, they were then combined to make the final verification decision.

3.5. Fusion

The ANN described above requires a total of $(232+1)*100 + (100+1)*16 = 23476$ parameters to train, and the classification hyper-planes are nonlinear. Having that many parameters to estimate caused a potential over-fitting problem for the ANN-based scoring system. We obtained no errors even for the 3-second systems in only a few iterations with only very limited training data. In the meantime, LDFs require only a total of $(232+1)*16 = 3728$ parameters to be trained, and the classification hyper-planes are linear. In contrast to the ANN, having fewer parameters prevented LDFs from modeling the data too well. These observations motivated us to develop theoretically well-defined schemes for fusing the scores from several verifiers, e.g., ANN and LDF.

In this study we fuse the two scores as shown in Figure 2. The two plots on the left are the distributions of the ANN and LDF confidence scores on some development set for target Mandarin, and the collectively non-target (a.k.a. imposter) languages, respectively. These plots make it clear that there is a language-specific threshold τ_j for each ANN or LDF classifier that minimizes C_{det} . If such thresholds *were* known, we could add it to all scores to yield the minimum possible C_{det} . By this score-shifting, not only the overall thresholds are set to 0 (hence the probability distributions are centered), but most importantly, C_{det} is minimized for each language.

These plots also reveal the fact that each classifier has a *bias* towards either more false-rejection or false-alarm errors, even when the minimum C_{det} is achieved. The ANN verifier is

biased towards making more false-rejection errors, while the LDF verifiers tend to produce more false-acceptance errors.

Furthermore, the ANN verifier has a longer error tail than the LDF verifier. Whenever these different two classifiers have a different error pattern, a well-adjusted fusion scheme provides us with the advantage of being able to remove such biases.

After computing these language-specific thresholds in each component we evaluate the fused confidence score as:

$$s_i^{Fused} = w_{1i}(s_i^{ANN} - \tau_i^{ANN}) + w_{2i}(s_i^{LDF} - \tau_i^{LDF}) \quad (2)$$

In our current study, we set both w_{1i} and w_{2i} equal to 1/2 for convenience. A performance improvement will be gained if these weights are trained as well.

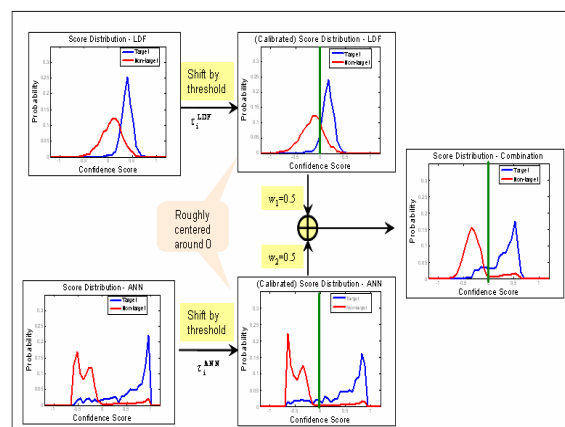


Figure 2: The probability distributions of the ANN and LDF confidence scores on development set. The score distributions are very different, and each verifier has its own bias towards making either more false-acceptance or more false-rejection errors.

4. Experiments

In this section, we report our experiments on the 30-second primary evaluation data set. Similar conclusions can be drawn from the 10- and 3-second primary evaluation sets and all the dialect evaluation sets.

4.1. Comparison of the Classifiers

In Figure 3 we compare the performance of ANN, LDF and the fusion systems on the 30-second primary evaluation set. With much less parameters, the LDF system obtained comparable result as ANN. The advantage of the fusion of ANN and LDF was clearly shown. With the decision strategy described in Section 3, we can compensate for the disadvantages of ANN and LDF, and obtain a better overall verifier.

4.2. Estimation of the Thresholds

In addition to minimizing C_{Det} for each individual classifier, centering the scores in individual verifiers makes it possible to safely combine scores from different systems. For this purpose, we constructed several validation sets from the 1996 and 2003 evaluation test sets to serve as development sets for threshold estimation. The composition of these validation sets is

summarized in Table 1. In particular, the data for target languages are the same but the ‘unknown’ language data differ, and were used to ensure proper handling of “other” languages. In order to set a specific language as ‘unknown’, we mask the corresponding output of ANN and LDF for that language. The ‘unknown’ language data are highlighted in Table 1 for clarity. The proportion of the target/unknown language data was chosen in a wide range to assure the threshold robustness.

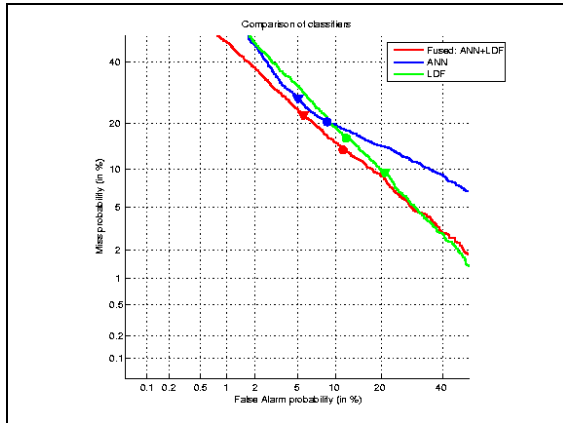


Figure 3: Performance comparison of the ANN, LDF and fusion systems on the 30-second primary evaluation set.

Table 1: Four validation sets are constructed to calculate the language-specific thresholds for each of the component classifiers, ANN and LDF. The validation sets have different amounts of ‘unknown’ language data to be able to handle the unknown 2005 evaluation data.

Set #1	Set #2	Set #3	Set #4
English03	English03	English03	English03
Hindi03	Hindi03	Hindi03	Hindi-03
Japanese96	Japanese96	Japanese96	Japanese96
Korean03	Korean03	Korean03	Korean03
Mandarin96	Mandarin96	Mandarin96	Mandarin96
Spanish03	Spanish03	Spanish03	Spanish03
Tamil03	Tamil03	Tamil03	Tamil03
Russian03	Russian03	Russian03	Russian03
Arabic03	French03	Vietnamese03	Arabic96
	German03	Farsi03	Farsi96
			French96
16.08% ‘unknown’ Language	22.33% ‘unknown’ Language	22.32% ‘unknown’ Language	27.71% ‘unknown’ Language

NIST’s Matlab-based DET curve plotting functions [15] can be used to find the language-specific thresholds τ_i for which the average of false-acceptance and false-rejection rates can be minimized. We calculated these thresholds τ_i for each of the 7 primary target languages and for each component ANN and LDF classifiers. This resulted in 4 threshold values per language, one for each validation set. We noticed that these thresholds did not show much variability even though the ‘unknown’ language data include different language data from different evaluation sets. For example, the 4 thresholds for the Mandarin language were found to be 0.2411, 0.2319, 0.2359

and 0.2389 in the LDF-based system and 0.0497, 0.1295, 0.0863 and 0.0913 in the ANN-based system. Taking the averages, we determined the thresholds τ to be 0.2369 and 0.0897 for the LDF and the ANN systems, respectively.

4.3. Robustness of Threshold Setting

In order to investigate the robustness issue of threshold setting further, we designed 4 more validation sets, in which there were 6.82%, 32.33%, 36.54% and 51.28% ‘unknown language’ data. Combined with the 4 validation sets described in Section 4.2, we can get the averaged thresholds for the first 4, 6 and the whole 8 validation sets. Using these averaged thresholds, 3 fusion systems were built for the NIST LRE05 30-second evaluation set. The performance summary is shown in Figure 4. We can see that the three DET curves were nearly overlapped with each other. It shows our threshold selection method is robust. This is due to the fact that the confidence scores defined in Eq. (1) can well model the separation of target and non-target languages.

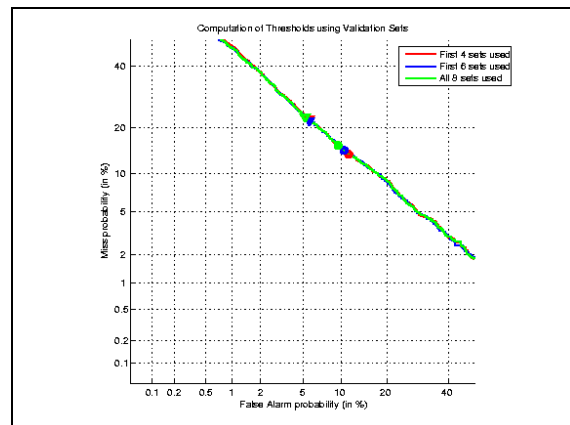


Figure 4: Performance of fusion systems with averaged thresholds obtained from 4, 6 and 8 validation sets.

4.4. High Confidence Trials

The usage of the confidence score, defined in Eqs. (1) and (2), was proven to be quite reasonable in high-confidence trials. Such trials were defined by limiting verification decisions using only 50% of the “true” decision trials with the highest confidence scores, and 50% of the “false” decision trials with the lowest scores. As shown in Figure 5, the performance of high-confidence trials achieved an equal error rate (EER) of about 4%, which was improved significantly over that of about an EER of 12% obtained with verification trials on all data.

In the 2005 NIST LRE systems, only two sites demonstrated such a performance improvement [16]. We believe that good high-confidence trials rely on a proper definition of confidence scores. For example, the value of the score difference, defined in Eq. (1), serves as a good indicator whether a correct classification decision can be made and how far it is from the decision boundary. Hence, a relatively large value usually implies a high confidence. It also shows that discriminative training of verification systems based on maximizing this difference, as in MCE training [9], is equivalent to minimizing the number of low-confidence trails,

and therefore enhancing the performance of systems focusing on high-confidence trials

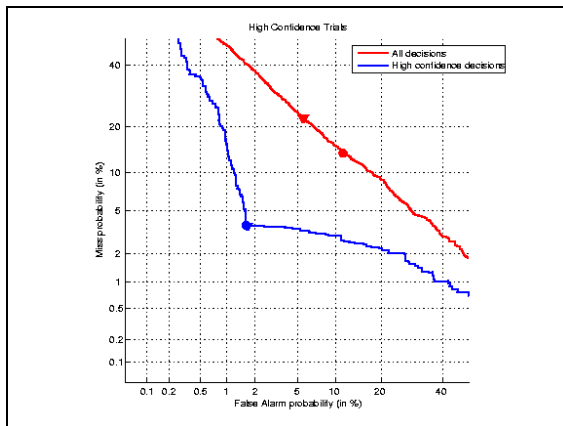


Figure 5: High-confidence trials for the 30-second primary evaluation set.

5. Summary and Future Work

We have presented the GT-IIR language recognition system designed for the 2005 NIST LRE. Instead of using the scores computed from the P-PRLM and BOS systems directly to make language recognition decisions, we used the scores from them for all competing languages to serve as input features to train the LDF and ANN verifiers, and fuse the output verification scores to make final decisions. Both the LDF and ANN classifiers can be obtained with discriminative training. For the LDF verifier with a small number of parameters we achieved a performance comparable with that of the ANN verifier, which is much more complex than the LDF verifier. We have also shown that the distribution of confidence scores from the ANN and LDF verifiers exhibited large diversity, which is ideal for score fusion. Experiments have demonstrated the fused system achieved a better performance than systems based on the individual LDF and ANN classifiers.

Discriminative classifier design also demonstrated a distinct advantage for systems based on high-confidence trials. By maximizing the separation between the models of the target and competing languages, the number of high-confidence trials is significantly increased and equivalently the performance of verification systems based on high-confidence trials is thus significantly improved.

The success of fusion in this study also encourages us to seek more verifiers with different error patterns from the ANN and LDF verifiers, and incorporate them into the current framework to reduce the overall verification error rates.

One potential area for improvement is to design language recognition systems that optimize multiple objectives for all the languages simultaneously. We have proposed an iterative constrained optimization algorithm that can be extended to meet the above requirements [17]. Another improvement is to move from maximizing the performance of a single operating point on the DET curves to optimizing the overall behavior of the DET curves. Recent studies on receiver operating characteristic (ROC) optimization, such as minimizing the area

under the ROC curve, have shown new promising directions. One such framework, called ensemble classifier design [18], is an ideal way to extend the success of learning single discriminative classifier to training multiple discriminative classifiers that can cover a wide range of operating conditions in order to meet new verification requirements.

6. References

- [1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no.1, pp. 31-44, 1996.
- [2] B. Ma, H. Li and C.-H. Lee, "An acoustic segment modeling approach to automatic language identification," *Proc. InterSpeech05*, Lisbon, Portugal, September 2005.
- [3] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition," *Proc. Eurospeech03*, pp. 1345-1348, 2003.
- [4] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language Recognition with Support Vector Machines," *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 41-44, 2004.
- [5] S. Katagiri and C.-H. Lee, "A New Hybrid Algorithm for Speech Recognition Based on HMM Segmentation and Discriminative Classification," *IEEE Trans. on Speech and Audio Proc.*, Vol. 1, No. 4, pp. 421-430, 1993.
- [6] K.-Y. Su and C.-H. Lee, "Speech Recognition using Weighted HMM and Subspace Projection Approaches," *IEEE Trans. on Speech and Audio Prociessng*, vol. 2, no. 1, pp. 69-79, 1994.
- [7] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification, 2nd edition*, John Wiley & Sons, 2001.
- [8] S. Haykin, *Neural Networks: A Comprehensive Foundation, 2nd edition*, Prentice Hall, 1998.
- [9] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.
- [10] NIST, "The 2005 NIST Language Recognition Evaluation Plan," 2005.
- [11] Language Identification Corpus of the Institute for Infocomm Research, <http://sdp.i2r.a-star.edu.sg>.
- [12] Linguistic Data Consortium (LDC), "The CallFriend Corpora".
- [13] H.-C. Wang, "MAT-a project to collect Mandarin speech data through networks in Taiwan," in: *Int. J. Comput. Linguistics Chinese Language Process.* 1 (2) (February 1997) 73-89.
- [14] <http://cslu.cse.ogi.edu/corpora/corpCurrent.html>
- [15] NIST, "DET-Curve Plotting Software for Use with MATLAB," http://www.nist.gov/speech/tools/DETware_v2.1.targz.htm
- [16] A. Martin and A. Le, "Part II: NIST Presentation of Results," *Record of 2005 NIST Language Recognition Evaluation Workshop*, December 2005.
- [17] S. Yaman and C.-H. Lee, "An Iterative Constrained Optimization Approach to Classifier Design," to appear in *Proc. ICASSP*, 2006.
- [18] S. Gao and C.-H. Lee, "An Ensemble Classifier Learning Approach to ROC Optimization," Submitted to *2006 ICPR*, Hong Kong, China, August 2006.