Development of An Approach to Automatic Language Identification based on Phone Recognition

Yonghong Yan Etienne Barnard Ronald A. Cole

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
P.O.Box 91000, Oregon 97291-1000, USA
Email: {yan, barnard, cole}@cse.ogi.edu, Fax: (503) 690-1334

Abstract

An Automatic Language Identification (LID) approach is presented. The baseline LID system consists of three parts: (1) Hidden Markov Models (HMM) based context independent phone recognizers, (2) language identification score generators and (3) a linear language classifier. The system exploits language dependent phonotactic constraints and prosodic information. Four methods are proposed to improve the system performance. Two bigram-based interpolated N-gram language models (forward and backward) are used to model the phone sequence constraints of different spoken languages. A context-dependent duration model interpolated by a context-independent duration model is used to capture the duration information. Comparison experiments between the linear classifier and neural-network based final classifiers were conducted. Finally, optimization of language model based on back propagation is proposed. The improved system was evaluated on an 11-language task, and performance reached 13.3% and 26.2% (error rate) for utterances averaging 45 seconds duration and 10 seconds duration, respectively. Compared with the baseline system performance, it shows the importance of the issues addressed in this paper for language identification.

1 Introduction

With the growth of global interaction, the demands for communications across the boundaries of languages are increasing. This challenges speech processing in at least one new respect: before the machine can understand the meaning of an utterance, it must identify which language is being used. The task for Language Identification (LID) is to quickly and accurately identify the language being used.

Theoretically, the differences between different spoken languages are manifold and large. Although these differences can be found at various levels (e.g. phoneme inventory, acoustic realization of phonemes, lexicon, phonotactics, etc.), how to reliably extract these features for reliable LID is still an unsolved problem.

Early approaches that exploited the acoustic feature vectors included a filter band based approach (Leonard and Doddington, 1974), LPC-based polynomial classification approach (Cimarusti and Ives, 1982), and formant vector quantization based approaches (Foil, 1986; Goodman, 1989). House and Neuburg (House and Neuburg, 1977) published the earliest language identification approach based on the different phonotactic constraints of different spoken languages. Later finite-state models were studied to capture the transition probabilities between broad phonetic categories (Li and Edwards, 1980). With the improvement of speech processing techniques, more sophisticated approaches have been studied recently. These include systems using vector quantization (Sugiyama, 1991), HMMs (Ueda and Nakagawa, 1990; Savic, Acosta and Gupta, 1991; Riek, Mistretta and Morgan, 1991; Nakagawa, Ueda and Seino, 1992; Zissman, 1993), neural-networks (Muthusamy and Cole, 1992), embedded keywords (Ramesh and Roe, 1994), syllabic spectral features (Li, 1994), the distinction between polyphonemes and monophonemes (Dalsgaard and Andersen, 1994), and subword recognition, exploiting acoustic/phonotactic constraints (Muthusamy, Berkling, Arai, Cole and Barnard, 1993; Tucker, Garey and Parris, 1994; Hazen and Zue, 1993; 1994; Reyes, Seino and Nakagawa, 1994; Lamel and Gauvain, 1994; Kadambe and Hi-

eronymus, 1994; Zissman and Singer, 1994; Berkling and Barnard, 1994).

These efforts have improved the performance of LID systems dramatically. Details can be found in a recent review paper (Muthusamy, Barnard and Cole, 1994). It was shown that accurate classification could not be achieved by simply using frame-based acoustic feature vectors; discriminaters such as phonotactic constraints are needed. Most state of the art LID systems directly or indirectly employ phonotactic information, modeled in terms of transitions among subword units. Generally, the performance of these systems is correlated with their complexity. A system that exploits several information sources outperforms one relying on only a single information source.

Although encouraging LID results have been reported in recent work mentioned above, further improvement requires the resolution of several issues, such as

1. How to balance the contradictory requirements between detailed modeling and database availability. Muthusamy *et al.* (1993) reported significant improvement by using phonemes rather than broad phonetic categories as the basis for LID, and others (e.g. Zissman *et al.*, 1994) have similarly found that more detailed modeling is beneficial. However, increasing the amount of detail in the models inevitably requires more training data for the models to achieve the same robustness.

2. How to model prosodic information.

3. How to best combine several information sources into one system.

4. How to extract the maximal useful information from the imperfect outputs of current subword recognizers. For potential telephone-oriented LID systems - which we study exclusively - this presents a severe problem. The phone recognition error rate is much higher for telephone speech than for high quality speech. How to improve speech recognition performance with the limited telephone bandwidth and how to handle channel noise are still challenging problems for speech recognition researchers.

In this paper, an approach to language identification based on phone recognition is presented. The system is composed of three parts: (1) an HMM based phone recognizer front end, (2) LID score generators and (3) a final classifier. It extends earlier approaches (Muthusamy *et al.*, 1993; Zissman *et al*, 1994) to address some of the issues raised above. A backward bigram based language model is proposed as an additional set of features to the conventional forward bigram. The introduction of this model improves the modeling accuracy (to capture both forward and backward information) without increasing the number of the total parameters in the model excessively. A natural extension of the duration model (a generalized context-dependent model interpolated with a context-independent model) is proposed. Experiments were conducted to evaluate the system performance by using a neural-network for information integration in comparison with a linear classifier. In order to compensate for limited phone accuracy, a method of optimizing the language identification models based on error back propagation is proposed.

The LID systems were evaluated on a six-language task and an 11-language task. The best results achieved were 8% and 17% (LID error rate) for 45-second long utterances and 10-second long utterances respectively on a standard six-language task, and 13% and 26% for 45-second long utterances and 10-second long utterances respectively on a standard 11-language task.

The rest of this paper is organized as: Section 2 presents the baseline system. Section 3 details the four methods proposed to enhance the system performance. In section 4, the database used in this study is described, and section 5 presents the evaluation of our baseline system. Section 6 contains the evaluation of the proposed improvements.

2 Baseline LID System

In this section, we will provide an overview of our baseline LID system. This system is designed to exploit language-dependent phonotactic information and duration information based on language-dependent phone recognition. This sys-

tem is used as both benchmark and platform for further development.

2.1 Models Used in the Baseline System

In this subsection, we will discuss the statistical models used for the baseline language identification system. Two sets of models are used to exploit sequential information and duration information of different languages.

2.1.1 Language Model

House *et al.* (1977) proposed that sequential constraints on phonemes could be exploited as an efficient approach to language identification. Their work showed that sequential constraints could be a powerful feature to distinguish different languages even when the speech events were described by broad-category classes. This idea has been used extensively in recent research (Muthusamy *et al.*, 1993; Hazen *et al.*, 1993; Zissman *et al.*, 1994; Kadambe *et al.*, 1994), and reflects the phonotactic differences between different languages (Ladefoged, 1993). Our previous work (Muthusamy *et al.*, 1993) showed that fine phone categories could lead to better performance than broad categories. Therefore, fine phone categories are used in this work.

The language model used in our baseline system is the commonly used bigram-based language model:

$$P_{LF} = \sum_{i=1}^{T} log(\alpha P(O_i|O_{i-1}) + \beta P(O_i)) \quad (1)$$

where O_i is the ith phone in the decoded best path, $P(O_i)$ is the unigram term and $\mathrm{P}(O_i|O_{i-1})$ is the bigram term; α and β are the weight coefficients. T is the total number of phones in the decoded utterance. This language model is based on the interpolated n-gram language model proposed in Jelinek (1990).

2.1.2 Duration Model

A duration model has been used in Hazen *et al.* (1993; 1994) to capture a certain class of prosodic information in different languages. In our studies, two

representations of duration distribution for each phone were evaluated initially, namely

- Gaussian densities and
- Histograms

Initial experiments showed the performance of these two representations to be similar. The histogram representation was selected for further work because of its computational simplicity. The duration model was estimated based on decoded phone strings in the training set.

2.2 System Architecture

Our baseline LID system is composed of three parts: (1) a language dependent phone recognizer, (2) LID score generators, and (3) a language classifier. The general architecture for an N-language task is given in Figure 1, where M is the number of recognizers used ($N \geq M$).

Perceptual experiments on a 10-language task (Muthusamy, Jain and Cole, 1994) found that the number of languages known was a significant factor in performance for human subjects. Similarly, several language-dependent recognizers (each with bigram models for each target language) enhance the performance of our automatic system. Similar system architectures were used in (Muthusamy *et al.*, 1993; Zissman *et al.*, 1994; Kadambe *et al.*, 1994).

2.2.1 Front End

The front end is composed of several general-purpose language-dependent phone recognizers. It takes the speech wave as input, performs short time LPC analysis and feeds the parameterized speech vectors into the recognizers. The output of the recognizer is the time-aligned phone string with an acoustic probability attached to each phone in the string.

For an N-language task, M ($N \geq M$) language-dependent phone recognizers run

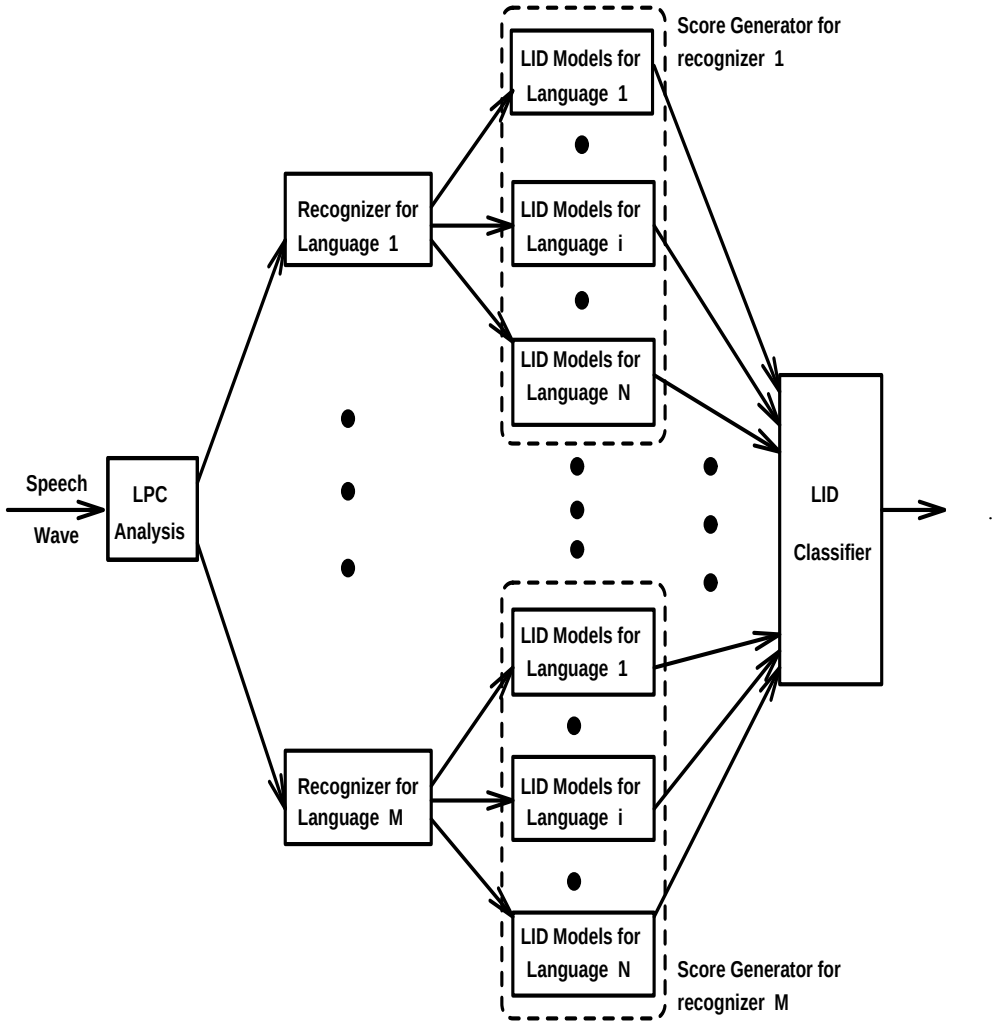

Figure 1: General Structure of the LID system For an N-language task with M recognizers

in parallel, and independently decode the input speech vectors into phone strings (Zissman, 1994).

In our implementation, six language-dependent phone recognizers were used ($M = 6$) since phonetic transcriptions in six languages (English, German, Hindi, Japanese, Mandarin Chinese and Spanish) were available.

We found the performance of the LID system to increase with an increasing number of recognizers in the system, consistent with the conclusion of the perceptual experiment. Of course, computational complexity is also increased with an increasing number of recognizers, which is a major drawback of this type of implementation.

2.2.2 LID Score Generator

As shown in Figure 1, the phone recognizer of each language has its own score generator in our system. Each score generator contains a set of LID models for each language in the task. The score generator takes the output (the decoded phone string) from each language-dependent recognizer, calculates the various LID feature scores, and provides them to the final classifier.

The phone recognizers were trained on the phonetically labeled data, and used to decode the data in the training set. The decoded results were used to train the language models and the duration models. By using each recognizer to decode the speech data in the training sets of all the languages in the task, the LID models for all the languages in the task can be estimated from the decoded phone strings in terms of the phone inventory of this specific recognizer.

During testing, after a recognizer decodes the input test utterance, the score generator for this recognizer will calculate the likelihoods of it being each language in the task separately. One advantage of this kind of implementation is that by averaging the likelihoods calculated by all the recognizers for one language, the bias created by different recognizers is reduced.

Before sending the scores (likelihoods) to the final language classifier, the lan-

guage modeling scores and duration scores are normalized by the number of phones in the best path decoded by their corresponding phone recognizers.

2.2.3 Final classifier

When multiple information sources are used, an important issue is how the sources should be combined to make a classification. In our baseline system, a linear classifier was trained to combine these information sources. An optimal linear classifier can be viewed as combining the probabilities with optimal weights.

3 Methods Used to Improve the Baseline System Performance

In this section, we present the four methods proposed in this paper to improve the performance of our baseline LID system. These methods were designed to cope with the four important issues for LID based on phone recognition discussed in the introduction.

3.1 Enhanced Language Model

The language model used in our baseline system exploits left context information; thus only the forward information is captured. Although a trigram-based language model can capture both the right and left context information, a larger database is needed in order to get a well-estimated model. Millions of words were used to train the trigram based language model for a single language in (Kadambe *el al.*, 1994). For a language with N phonemes, on the order of N^2 parameters need to be estimated for a bigram model, while for a trigram the number is order N^3. As a compromise, we propose using two bigram models $\mathrm{P}(O_i|O_{i-1})$ (forward) and $\mathrm{P}(O_i|O_{i+1})$ (backward) in the language model (Yan and Barnard, 1995a). One possible way to combine these two bigram models

into a language model is as:

$$P_{LFB} = \sum_{i=1}^{T} log(\alpha P(O_i|O_{i-1}) + \beta P(O_i|O_{i+1}) + \gamma P(O_i)) \quad (2)$$

Adding the backward bigram term enables the language model to capture both the right and left context information without adding too many parameters to be estimated. The contradictory requirements of detailed modeling and data efficiency is thus partially solved for the modeling of the phonotactic constraints.

3.2 Enhanced Duration Model

Since the variation in the durations of a phoneme in different contexts can be quite large, context-dependent duration modeling is desired. In order to decrease the number of parameters in the model, a duration model based on the generalized left context is proposed. For each phone, six duration models are estimated depending on whether its proceeding phone is a *vowel, fricative, stop, nasal, affricate or glide*. The duration models used in this paper are given as:

$$P_D = \sum_{i=1}^{T} log((1-\alpha)P(O_i|O_{i-1} \in S) + \alpha P(O_i)) \quad (3)$$

where $P(O_i|O_{i-1} \in S)$ is the context-dependent model, and S is one of the six broad categories. $P(O_i)$ is the original context independent monophone duration model, which is used here as a smoothing factor with weight α. In our experiment, α is set to 0.1.

3.3 Neural-Network Classifier

Recent research (Hazen *el al.*,1994) has shown that simply combining all the information sources with different weights may not result in the best system performance. Our pilot experiments also showed that the assumption of the independence of these information sources was not appropriate. We proposed to view the combination of scores as a classification problem rather than a combination of independent probabilities. A well-trained linear classifier can be viewed

as combining the scores with different optimal weights to give the best guess, but is known to be inferior to non-linear classifier in many circumstances. In our experiment, neural-network based pattern recognizers were studied as an alternative final classifier.

A feed-forward neural network with one hidden layer and full connections between successive layers is used to learn the relations among these scores in our system. The output of the neural network is the final LID result of the input utterance. The neural network was trained with conjugate gradient optimization (Barnard and Cole, 1989).

3.4 Optimization of the Language Model

For free vocabulary continuous telephone speech, our speaker-independent phone recognizers are around 45% to 55% accurate. This reduces the accuracy of our system, since approximately 50% of phones in the decoded path contain erroneous information about the phonotactic constraints of the language.

In this section, we will discuss our attempt to optimize the language model, so as to improve robustness with respect to these errors.

Our approaches uses a cost function based on the maximum likelihood LID scores calculated by the score generator (Yan and Barnard, 1995b). For any score generator in the system, the cost function for a particular input utterance is defined as:

$$E = (S_T - S_M)^2 \tag{4}$$

where

$$S_M = \max_i S_i, \quad i = 1, ..., N \tag{5}$$

S_i is the LID score for language i in an N-language task, S_T is the score for the expected target language($S_T \in S_i$), where:

$$S_i = \begin{cases} P_{LF}, & \text{for forward language model} \\ P_{LB}, & \text{for backward language model} \\ P_D, & \text{for duration model} \end{cases}$$

When $T = M$, *i.e.*, the correct language has the highest score, $E = 0$,
and when $T \neq M$, E provides a quantitative measurement for the error in the language model. By back propagating the error, optimization of the language model for this score generator will be achieved.

Here we give an example to illustrate how the optimization is realized; this example is for the optimization of the bigram term in the forward language model. Other terms in the language models can be optimized similarly.

By using (1) and (4), for the forward language model,

$$E = (S_T - S_M)^2 \tag{6}$$

$$= (\sum_{i=1}^{T} log(\alpha P(O_i|O_{i-1}) + \beta P(O_i)) - S_M)^2 \tag{7}$$

$$\frac{\partial E}{\partial P(O_i|O_{i-1})} = 2(S_T - S_M)\frac{\partial S_T}{\partial P(O_i|O_{i-1})} \tag{8}$$

$$= 2(S_T - S_M)\frac{\alpha}{\alpha P(O_i|O_{i-1}) + \beta P(O_i)} \tag{9}$$

We thus update $P(O_i|O_{i-1})$ by

$$\Delta P(O_i|O_{i-1}) = -\eta \frac{\partial E}{\partial P(O_i|O_{i-1})} \tag{10}$$

where η is the learning rate. *i.e.*,

$$P_{new}(O_i|O_{i-1}) = P(O_i|O_{i-1}) + \Delta P(O_i|O_{i-1}) \tag{11}$$

After one iteration, the re-normalization of the bigram probability results in a new bigram term for the language model. The training data for estimating the

language model are divided into two sets, *set 1* and *set 2*. The ratio of the sizes of these two sets is roughly 2:1. *Set 1* is used to derive the original bigram by the traditional linear operation. Both sets are used for the optimization. Cross validation on the development set is performed at each iteration by the score generator. The performance (LID correct rate) is used as the stopping criterion. The advantage of this back propagation based optimization is that it dynamically enhances the discrimination among the languages. The resulting bigram is no longer a simple function of the training data. This compensates for the problem caused by limited training data to some extent.

With this process, the modeling of phone sequences which are not sufficiently represented in the training data can be enhanced. The cross validation procedure and the gradient-based descent guarantee that the optimization procedures will modify probabilities to improve the discrimination between the languages. The actual optimization steps are illustrated in Figure 2. An important issue in using this optimization method is how to prevent the model from being too dependent on the data set, that is, how to prevent over-fitting is crucial to the success of the procedure. In our implementation, three methods to prevent over-fitting were used: (1) The diversities of different data sets were utilized (Three independent sets were used. One was the original training set for language models, the second set was used together with the original training set for optimization. The third set was used for the cross validation and training was stopped based on this.). (2) Instead of updating the bigram after each data presentation, batch mode was used (For each term, the ΔP needed were accumulated. The actual updating took place after all the utterances were processed once.). (3) A small learning rate was used.

4 Database and Method

The database used in this study is from the OGI_TS database (Muthusamy, Cole and Oshika, 1992), which is a multi-language telephone speech database designed

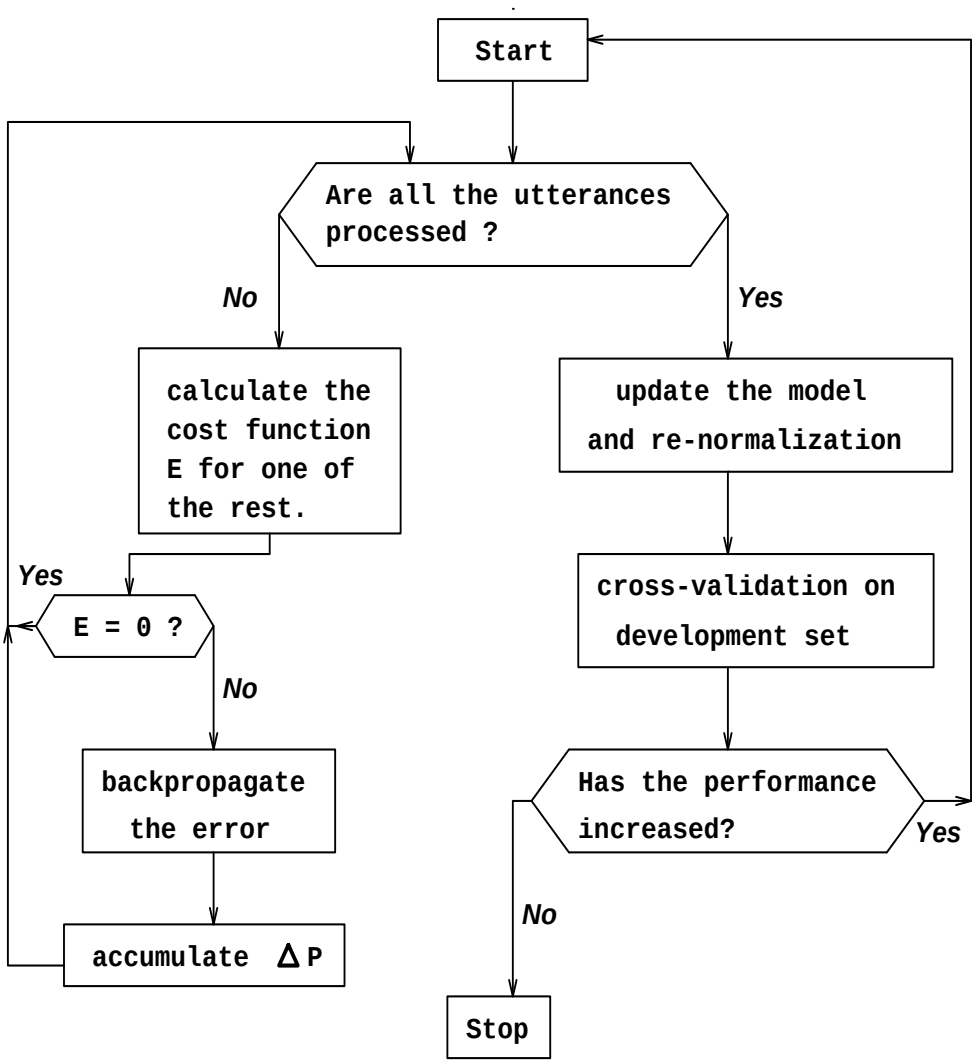

Figure 2: Optimization of the LID model based on back-propagation

for language identification research. Currently there are speech data from 11 languages (English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese) in the database. Six of them (English, German, Japanese, Mandarin and Spanish) have been phonetically transcribed. Since the training of our phone recognizer needs phonetically transcribed data, the maximum number of recognizers available within this database is six ($M = 6$). We evaluated our approach with two tasks, namely on only those languages which were transcribed ($N = 6$), and on all languages ($N = 11$).

4.1 Data for the Six-language Task

Our first set of experiments is with the six languages which have been transcribed phonetically.

Two parts of utterances in the database were used: "story-before-the-tone" (story-bt) and "story-after-the-tone "(story-at). The training set contained 777 utterances total, and was divided into the following three sets.

- Training set 1. There were 300 utterances in this subset; all were phonetically labeled.
- Training set 2. There were 400 utterances in this subset; most of them were not labeled.
- Development set. There were 77 utterances in this subset; most of them were phonetically labeled.

In each set, the numbers of utterances in each language were approximately balanced.

The test set used was the same test set used by the National Institute of Standards and Technology (NIST) in their March 1994 evaluation for these six languages. It contained 112 whole utterances (45-second long) and 370 ten-second utterances, which were also part of OGI_TS database. None of the above four sets

overlapped.

Training set 1 was used to train the language-dependent phone recognizers for each of these six languages. Training set 2 was used to train the LID models used in this study. The development set was used to evaluate the performance of the recognizers and all these sets were used to train the final classifier.

4.2 Data for the 11-language Task

For the 11-language task, four parts of the utterances in the database were used: "story-before-the-tone" (story-bt),"story-after-the-tone" (story-at), "rooms" (room) and "numbers" (num). The training set had 1785 utterances total (data used in the six-language task were included), which was further divided into following three sets.

- Training set. There were 716 utterances in this subset.
- Development set 1. There were 631 utterances in this subset.
- Development set 2. There were 438 utterances in this subset.

Again, the numbers of utterances in each language were approximately balanced. The test set for the 11-language task used was the test set used by NIST in the March 1994 evaluation. It contained 195 whole utterances and 625 ten-second utterances.

None of the above four sets overlapped.

The training set was used to train the LID models for the system, the training set and the development set 1 were used to perform the optimization while development set 2 was used for cross validation. All three of these sets were used to train the final classifier.

5 Baseline System Evaluation

Our baseline system was evaluated on the two tasks using the database described before. These results provided a benchmark for the evaluation of our further re-

search.

5.1 Implementation of the Phone Recognizers

Six continuous HMM based phone recognizers were implemented using the phonetically labeled data. Each phone was modeled as a three-state left-to-right HMM model. Four Gaussian mixtures were used to model the probability density for each state in a model. Speech wave data were parameterized every 20ms with 10ms overlap between contiguous frames. For each frame a 26-dimensional feature vector was calculated: 12 LPC_cepstra, 12 delta cepstra, normalized energy and delta energy.

(Table 1 here)

The recognizers were trained using training set 1 described above and phone recognition accuracies were evaluated on the labeled data in the development set. Table 1 gives the number of phones used in each language and the corresponding phone recognition accuracy.

5.2 Benchmark Evaluation

After the phone recognizers were implemented, they were used to decode the data. The decoded results were used to train the language models and the duration models described in the baseline system section of this paper.

Table 1: Number of Phones and its Recognition Accuracy of Each recognizer

The size(number of phones) and phone recognition accuracy for each recognizer.						
	English	German	Hindi	Japanese	Mandarin	Spanish
Size	41	41	45	26	41	30
Accuracy	46.8%	46.6%	48.1%	56.3%	36.3%	54.6%

The final classifier was trained on both the training data and the development data.

The following experiments were performed:

1. Experiment with the forward bigram based language model (F)
2. Experiment with the duration model (D)
3. Experiment with the complete baseline system (BS)

The best results achieved are given in Table 2 and Table 3.

(Table 2 here)

(Table 3 here)

The results of the first two experiment provide the quantitative measurement of the merit of each component for LID. The third experiment gives the benchmark performance of our baseline system when both LID models are used.

Table 2: Baseline System Evaluation Results I.

Benchmark results (error rate) on the six-language task.			
	F	D	BS
45-second	15.2%	44.6%	13.4%
10-second utterance	26.0%	54.9%	23.8%

Table 3: Baseline System Evaluation Results II.

Benchmark results (error rate) on the 11-language task.			
	F	D	BS
45-second	23.6%	58.0%	21.5%
10-second utterance	34.9%	68.0%	33.0%

6 Experiments with the Proposed Improvement Methods

In this section, we will present our experiments with the improvement methods proposed in this paper. The front end (phone recognizers) developed for the baseline system was used throughout. First we present the results of the final complete system when all the enhanced methods were used. Then we will present the comparison between the baseline system and the baseline system enhanced by each individual improvement method to illustrate how the improvement was achieved.

6.1 Experiments with the Complete Final System

These experiments are similar to those presented in the baseline evaluation section. Two experiments were conducted with the Final System (FS), with all the methods applied to the baseline system. The front end (the six language dependent recognizers) is the same as the front end in our baseline system. The score generator has three sets of LID models: the forward bigram based language model, the backward bigram based language model and the enhanced duration model. The final classifier is a neural network. The data sets used for training, developing and testing were the same as those used in the baseline system evaluation. All the results of the above experiments are given in Table 4. For comparison, the performance of baseline system is also listed.

(Table 4 here)

Table 4: Language Identification Correct Rate With Six-language Task

System performance : baseline vs. final system

_6: six-language task; _11: 11-language task

	BS_6	BS_11	FS_6	FS_11
45-second	13.4%	21.5%	9.0%	13.3%
10-second	23.8%	33.0%	18.4%	26.2%

The confusion matrices for the best system on the 11-language task are given in Table 5 (45-second long utterance) and Table 6 (10-second long utterance). The rows of the matrices correspond to the languages actually being spoken and the columns indicate the languages identified.

(Table 5 here)

(Table 6 here)

Note that the confusions are not only between linguistically similar languages; some languages (such as Hindi and Spanish) are consistently chosen more often. One possible reason may be the language-dependent phone recognizer, which uses a forced-choice best-path search. The approximation of the underlying linguistic input of an utterance in one language by the phone inventory of another language decreases the linguistic distinction between different languages. Other reasons for confusion include imperfect performance of the recognizer, lack of an explicit model of acoustic information and not strictly balanced sizes of training data for languages.

Table 5: Confusion matrix for the 45-second long utterances

Confusion Matrix for Our Best Approach for the 11-language Task Language name is denoted by its first two characters.											
Language	EN	FA	FR	GE	HI	JA	KO	MA	SP	TA	VI
EN	17	0	1	0	0	0	0	0	1	0	0
FA	0	17	0	2	0	0	0	0	0	0	0
FR	0	1	15	1	0	0	1	0	0	0	0
GE	0	0	0	18	0	0	0	0	0	2	0
HI	0	1	1	0	17	0	0	0	1	0	0
JA	0	0	0	0	1	17	1	0	0	0	0
KO	0	0	0	0	2	0	15	0	0	0	0
MA	0	0	1	1	0	1	1	13	0	0	0
SP	0	0	1	0	1	0	0	0	14	0	1
TA	0	0	0	0	0	0	0	0	0	14	0
VI	0	0	0	0	2	0	1	0	0	0	12

Table 6: Confusion matrix for the 10-second long utterances

Confusion Matrix for Our Best Approach for the 11-language Task Language name is denoted by its first two characters.											
Language	EN	FA	FR	GE	HI	JA	KO	MA	SP	TA	VI
EN	59	0	1	0	0	2	3	0	0	4	0
FA	1	45	3	1	3	0	0	2	1	1	1
FR	1	3	36	4	5	0	3	4	6	0	0
GE	2	2	0	52	3	0	0	0	2	2	2
HI	3	3	1	2	45	2	3	3	1	1	1
JA	0	0	2	0	7	43	1	1	5	1	1
KO	1	4	1	0	2	2	31	1	1	1	1
MA	0	0	3	1	2	3	5	37	1	0	0
SP	1	0	1	1	8	4	0	0	41	1	1
TA	0	0	0	0	2	0	0	0	1	40	0
VI	2	0	2	1	4	0	3	0	3	0	32

6.2 Experiments with Each Method

In order to compare and illustrate the impact of each method proposed on the baseline system, the following experiments were conducted.

1. Baseline system plus backward bigram based language model (BS_B)
   In this experiment, in addition to the two models used in the baseline system, the backward bigram based LID model was used as the third set of models to enhance the modeling accuracy. The data used to train the new model were the same as those used to train the language model in the baseline system.

2. Baseline system with the enhanced duration model (BS_ED)
   In this experiment, the duration model in the baseline system was replaced by the enhanced duration model. The new duration model was trained with the data used to train the baseline duration model.

3. Baseline system with the neural network as the final classifier (BS_NN)
   In this experiment, the linear classifier was replaced by a feed forward network. The data used to train the linear classifier were used to train the network. Various neural network architectures (different numbers of hidden nodes) were tested on the development set. The best architecture (with 20 hidden nodes) was used to do the final training and testing.

4. Baseline system with optimization (BS_OP)
   In this experiment, the forward language model was optimized by the formula given in (11). Both the training data and the development data were used in the optimization. Development set 1 was used in the optimization procedure with the training set, and development set 2 was used for cross validation. The optimization procedure stopped after three iterations, with the learning rate(η) set to 0.001.

The results of above experiments are given in Table 7 (six-language task) and Table 8. Figure 3 and Figure 4 illustrate the effect of optimization on the bigram terms in the language model and the language recognition performance. In Figure 3, only the changes of the probabilities of a few bigram terms are shown. These are in the models of the score generator associated with the English phone recognizer. Clearly, bigram values are only altered by a small amount. Nevertheless, the optimization procedure improves the performance of the score generator consistently, and different score generators improve their performance on different languages in the task. As shown in Figure 4, for the English Score generator, the performance on data from English, Hindi, Japanese and Spanish are improved by the optimization.

(Figure 3 here)

(Figure 4 here)

(Table 7 here)

(Table 8 here)

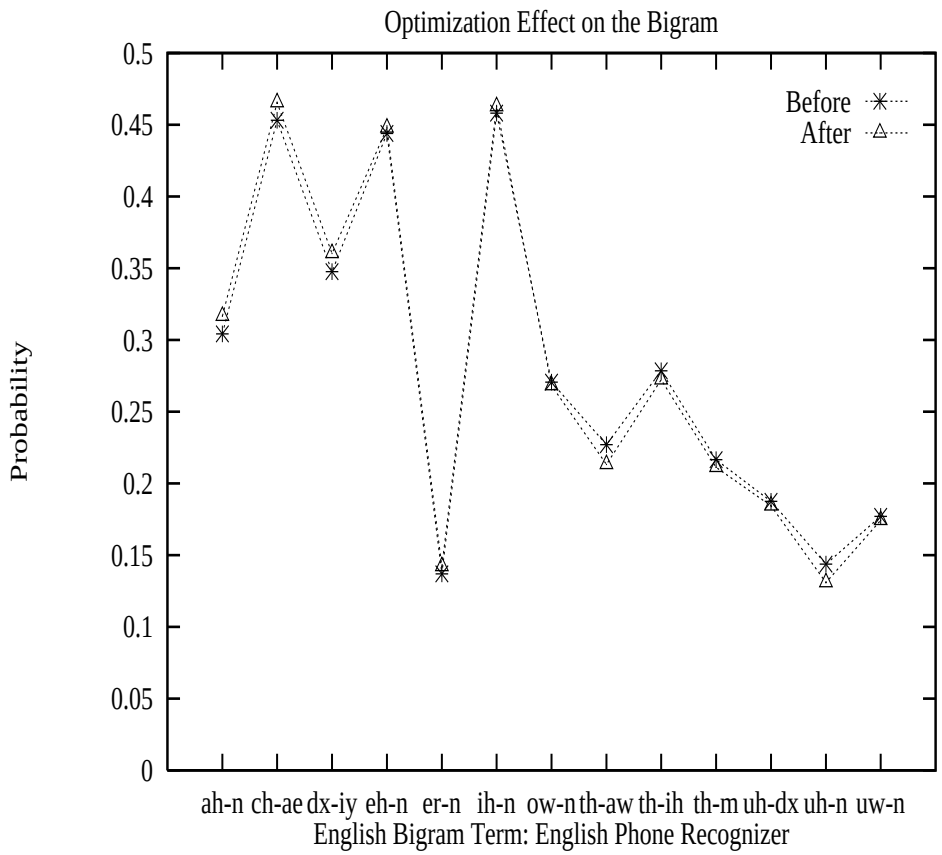

Figure 3: Analysis of Language Model (English Front End)

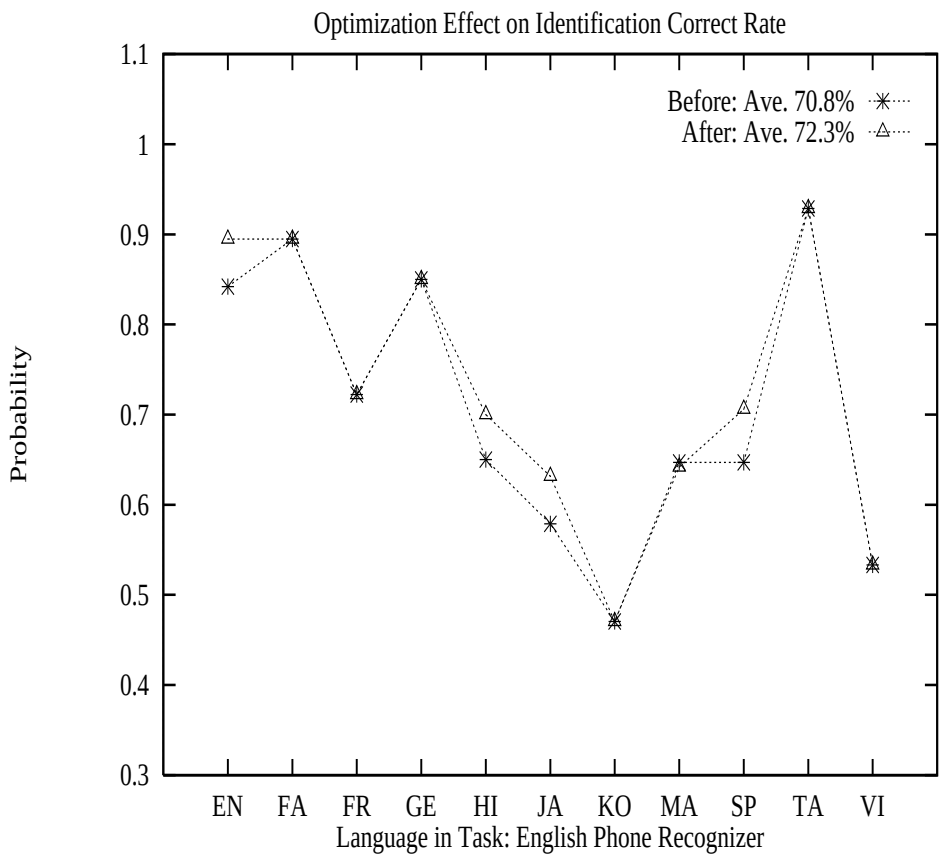

Figure 4: Optimization Results (English Front End)

Table 7: Language Identification Error Rate With Six-language Task

The performance for 45-second long utterances and 10-second long utterances.				
	BS_B	BS_ED	BS_NN	BS_OP
45-second	11.6%	12.5%	10.7%	10.7%
10-second	21.1%	23.8%	20.5%	20.3%

Table 8: Language Identification Error Rate With 11-language Task

The performance for 45-second long utterances and 10-second long utterances.				
	BS_B	BS_ED	BS_NN	BS_OP
45-second	17.4%	21.0%	16.4%	16.9%
10-second	29.9%	32.0%	28.3%	28.6%

7 Discussion

Encouraging language identification results have been achieved with the approach based on language-dependent phone recognition presented in this paper. The improved baseline system compared favorably with previously reported results on the same 6- and 11-language tasks.

The evaluation of the different LID models on the same baseline system platform with the same data sets provides a general understanding of the relative importance of the LID models when they are used alone. The results show that when used alone, the forward bigram based language model plays the most important role in our system.

To address the contradictory requirement between detailed modeling and data availability, we proposed the backward bigram based language model as an additional set of language model to the forward bigram based language model. Without increasing the requirement of training data, it provides a good approximation of the triphone based language model. We achieved approximately 20% error reduction on the 45-second utterance part on both tasks, and 10% error reduction on the 10-second long utterance part compared to the baseline system. The introduction of the new duration model did not yield a great improvement. Although it helps the system slightly, how to model the prosodic information effectively is still an interesting issue for future research. The comparison of the neural-network classifier and the linear classifier showed that non-linear combination of these information sources is necessary for language classification. Using the neural-network results in more than 20% error reduction on the 45-second utterance tasks and 15% error reduction on the 10-second tasks compared with the linear classifier.

The proposed optimization method demonstrates the promise and feasibility of improving the LID accuracy by increasing the discrimination of the language model using an automatic optimization procedure. With this back-propagation based optimization procedure, the bias created by the poor linguistic coverage in

the training data and the poor performance of the phone recognizers is alleviated. We achieved approximately 20% error reduction on the 45-second utterance tasks and 15% error reduction on the 10-second utterance tasks.

When all the proposed methods are applied to the baseline system, on the six-language task, we achieved approximately 35% error reduction on the 45-second utterances and more than 20% error reduction on the 10-second utterances. While on the 11-language task, more than 40% error reduction on the 45-second utterances and more than 25% error reduction on the 10-second utterances was achieved, which shows that our methods for improving the system performance is generalizable to a larger task.

We achieved more improvement on the 45-second utterance tasks than on the 10-second utterance tasks. One major reason is that all our LID models are statistical models based on the decoded phone strings (so the acoustic information is not exploited directly). On average the 45-second utterance segments are four times as long as the ten-second segments, so the possible linguistic bias (phone occurrence and sequential phone coverage) for long utterances is much smaller than for the short utterances. Further improvement on ten-second utterances may thus require using the differences between languages at the acoustic level: when phone constraints are not well reflected in short utterances, acoustic information may become a dominant feature. We are studying methods of incorporating acoustic information directly into our current LID system without requiring transcribed training data in all the languages.

reference

Barnard, E. & Cole, R.(1989). A nural-net training program based on conjugate-gradient optimization. *Technical Report CSE 89-014*, Oregon Graduate Institute.

Berkling, K. & Barnard, E.(1994). Language identification of six languages based on a common set of broad phonemes. In *Proceedings of International Conference on Spoken Language Processing*, pp.1891-1894.

Cimarusti, D. & Ives, R.B. (1982). Development of an automatic identification system of spoken languages: phase I. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.1661-1663.

Dalsgaard, P. & Andersen, O.(1994). Application of Inter-language phoneme similarities for language identification. In *Proceedings of International Conference on Spoken Language Processing*, pp.1903-1906.

Foil, J.T.(1986). Language identification using noisy speech. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.861-864.

Goodman, F.J., Martin, A.F. & Wohlford, R.E.(1989). In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.528-531.

Hazen, T.J. & Zue, V.W.(1993). Automatic language identification using a segment-based approach. In *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1303-1306.

Hazen, T.J. & Zue, V.W.(1994). Recent improvements in an approach to segment-based Automatic language identification. In *Proceedings of International Conference on Spoken Language Processing*, pp.1883-1886.

House, A.S, & Neuburg, E.P.(1977). Toward automatic identification of the language on an utterance.I.preliminary methodological considerations. *Journal of Acoustic Society of America* 62(3), 708-713.

Jelinek, F.(1990) Self-organized language modeling for speech recognition. In *Readings in speech recognition*(A. Waibel & K.F. Lee, ed.), pp.450-506. Morgan Kaufmann, Palo Alto, CA.

Kadambe, S. & Hieronymus, J.L.(1994). Spontaneous speech language identification with a knowledge of linguistics. In *Proceedings of International Conference on Spoken Language Processing*, pp.1879-1882.

Ladefoged, P.(1993). *A course in phonetics*, Third Edition, pp.266-272. Harcourt Brace Jovanovich, Inc.

Lamel, L.F. & Gauvain, J.L.(1994) Language identification using phone-based acoustic likelihoods. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp. 293-296.

Leonard, R.G. & Doddington, G.R.(1974). Automatic language identification. *Technical Report RADC-TR-74-200/TI-347650*, RADC/Texas Instruments, Inc, Dallas, TX.

Li, K.P. & Edwards, T.J. (1980). Statistical models for automatic language identification. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.884-887.

Li, K.P.(1994). Automatic language identification using syllabic spectral features. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.297-300.

Muthusamy, Y.K, Cole, R.A. & Oshika, B.T.(1992). The OGI multi-language telephone speech corpus. In *Proceedings of International Conference on Spoken Language Processing*, pp.895-897.

Muthusamy, Y.K & Cole, R.A.(1992). Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings of International Conference on Spoken Language Processing*, pp.1007-1010.

Muthusamy, Y.K., Berkling, k., Arai, T., Cole, R.A & Barnard, E.(1993). A comparison of approaches to automatic language identification using telephone speech. In *Proceedings of the European Conference on Speech Communication and Technology*, pp.1307-1310.

Muthusamy, Y.K., Jain, N. & Cole, R.A.(1994). Perceptual benchmarks for automatic language identification. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.333-336.

Muthusamy, Y.K., Barnard, E. & Cole, R.A.(1994). Reviewing Automatic Language Identification. In *IEEE Signal Processing Magazine*, pp.33-41. October 1994.

Nakagawa, S., Ueda, Y. & Seino, T.(1992). Speaker-independent, text independent language identification by HMM. In *Proceedings of International Conference on Spoken Language Processing*, pp.1011-1014.

Ramesh, P. & Roe, D.B.(1994). Language identification with embedded word models. In *Proceedings of International Conference on Spoken Language Processing*, pp.1887-1890.

Reyes, A.A., Seino, T. & Nakagawa, S.(1994). Three language identification methods based on HMMs. In *Proceedings of International Conference on Spoken Language Processing*, pp.1895-1898.

Riek, L., Mistretta, W. & Morgan,D.(1991). Experiments in language identification. *Technical Report SPCOT-91-002*, Lockheed Sanders, Inc., Nashua, NH.

Savic, M., Acosta, E. & Gupta, S.K.(1991). An automatic language identification system. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.871-820.

Sugiyama, M.(1991). Automatic language recognition using acoustic features. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.813-816.

Tucker, R.C.F., Carey, M.J. & Parris, E.S.(1994). Automatic language identification using sub-word models. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.301-304.

Ueda, Y. & Nakagawa, S.(1990). Prediction for phoneme/syllable/word category and identification of language using HMM. In *Proceedings of International Conference on Spoken Language Processing*, pp.1209-1212.

Yan, Y. & Barnard, E. (1995a). An Approach to Automatic Language Identification based on Language-dependent Phone Recognition. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.3511-3514.

Yan, Y. & Barnard, E. (1995b). Recent Improvements to a Phonotactic Approach to Language Identification. In *Proceedings of the Fifteenth Annual Speech Research Symposium XV*. pp.212-219.

Zissman, M.A.(1993). Automatic language identification using gaussian mixture and hidden markov models. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.399-402.

Zissman, M.A. & Singer, E.(1994). Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. In *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp.305-308.