Development of An Approach to Language Identification based on Language-dependent Phone Recognition

Yonghong Yan
B.E., Tsinghua University, P.R.China 1990

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science & Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

October 1995

The dissertation "Development of An Approach to Language Identification based on Language-dependent Phone Recognition" by Yonghong Yan has been examined and approved by the following Examination Committee:

Etienne Barnard
Associate Professor
Thesis Research Adviser

Ronald A. Cole
Professor

Mark Fanty
Assistant Professor

Wayne Ward
Research Scientist

Pieter Vermeulen
Senior Research Associate

Dedication

To my parents, Ruiying and Shien, my wife, Xu

Acknowledgements

I wish to express my foremost thanks to my advisor, Dr. Etienne Barnard, who worked closely with me during past years. His valuable suggestions and ideas greatly influenced the direction of this dissertation. In fact, he is the co-author of this work. Also I would like to express my deepest gratitude to Prof. Ronald A. Cole for all the advice, support and help. Without their valuable insight and discussion, this work could not be finished at this moment.

I would like to thank Dr. Mark Fanty, Dr. Wayne Ward (CMU) and Dr. Pieter V Vermeulen for the reviewing of this work and suggestions in shaping the thesis. Also, I am indebted to other members in CSLU (Center for Spoken Language Understanding) for all the helps during the past years. Specially I would like to thank all the CSLU members for the tolerance of CPU time and disk space I used.

Finally, I would like to thank my wife, Xu, for the consistent support.

Contents

List of Tables

List of Figures

Abstract

Development of An Approach to Language Identification based on Language-dependent Phone Recognition

Yonghong Yan, Ph.D.
Oregon Graduate Institute of Science & Technology, 1995

Supervising Professor: Etienne Barnard

The goal of Language Identification (**LID**) is to quickly and accurately identify the language being spoken. Although the differences among different (spoken) languages are generally large by any sensible measure, automatic language identification remains a major challenge (perhaps indicating the immaturity of the field of speech processing).

Current language identification systems vary greatly in terms of information utilization and system complexity. Understanding all of these approaches in a unified framework is one of the major challenges in automatic language identification. In this dissertation we provide a partial unification by studying the roles of acoustic, phonotactic and prosodic information in a particular system for language identification.

A comparative study was first conducted on a common two-language task (English and Japanese) to get a grasp of these issues. The results from the comparative experiments were used as basis for the development of a general purpose language-identification baseline system.

Within this frame work, two novel LID information sources (backward language model and a context-dependent duration model) were introduced. These two models increased language modeling accuracy at a moderate cost in terms of training data. Also, a novel optimization method was introduced to enhance the discrimination between different languages. These methods led to substantial improvements in system performance. Preliminary studies into channel normalization, conversational speech and system adaptation to new languages were also pursued.

A general purpose LID software tool kit was developed based on the algorithm developed in this thesis work. The final LID system developed attained correct rates of 91% (45-second segments) and 77% (ten-second segments) on a commonly used nine-language task. This is one of the best results reported to date on these tasks.

Chapter 1

Introduction

With the growth of global partnership, the demands for communications cross the boundaries of languages are increasing. This gives rise to new challenges for automatic speech recognition: before the machine can understand the meaning of the utterance, it must identify which language is being used. The task of Automatic Language Identification (LID) is to quickly and accurately identify the language being used.

The applications of language identification include human and automatic translation services, emergency services, and multi-lingual information retrieval. Figure 1.1 shows, for example, how two people who speak different languages could communicate with each other via a multi-lingual spoken language system.

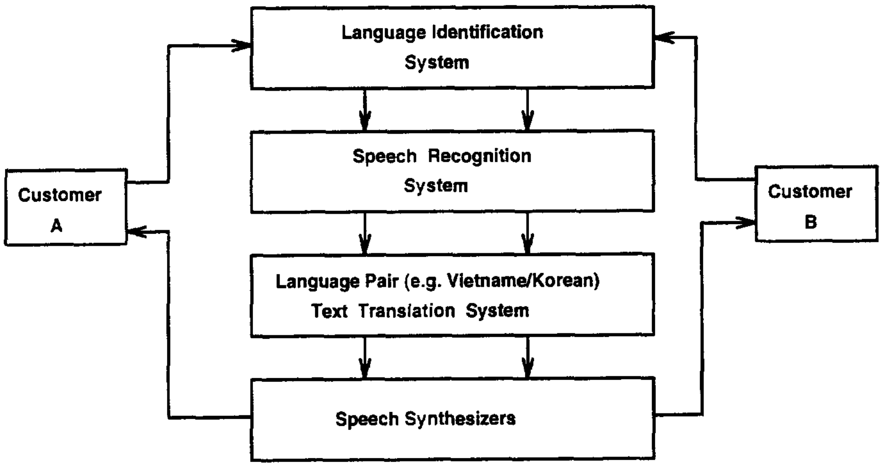

Language Identification System
Speech Recognition System
Language Pair (e.g. Vietname/Korean) Text Translation System
Speech Synthesizers
Customer A
Customer B

Figure 1.1: Communication using different languages via machine aid

Compared with other areas in speech processing, Automatic Language Identification is a relatively new pursuit. Although it is similar to Automatic Speech Recognition, Automatic

Speaker Identification and Accent Detection in some aspects, the differences between all of these tasks are substantial. The performance of a language-identification system is currently limited by several unsolved problems, such as:

1. What are the best features for language identification.
2. How to reliably extract these features for reliable LID.
3. How to model these features with limited data.
4. How to increase discrimination between different languages.
5. How to integrate multiple information sources into a unified framework.

In this dissertation, we describe the development of an approach to language identification based on language-dependent phone recognition, which shows our attempt to address these issues.

1.1 Background

In this section, we first discuss the inherent differences among different (spoken) languages. Based upon an understanding of these differences, we discuss the possible information sources that can be utilized in language identification and the difficulties in using these information sources.

1.1.1 Nature of the Problem

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, phonotactic regularities, and prosodics etc.), reliable language identification is still a challenge since reliable algorithms do not exist at any of these levels.

Composition of a spoken language: the Foundation for LID

The sounds of a spoken language can be described in terms of a set of abstract linguistic units called phonemes. A phoneme is the smallest contrastive unit in the phonology of a language. Each phone (the realization of a phoneme) has its unique articulatory configuration of the vocal tract. Different combinations of phonemes constitute different words. So different words consist of different sequences of phoneme strings that correspond to the vocal tract movements needed to produce the words. Furthermore, different combinations of words produce an infinite number of sentences that carry all the information that one wants to convey.

So a spoken language is composed of an infinite number of combinations of the set of phonemes; this is attached to various other sources of acoustic information, such as duration, pitch variation and stress.

Articulatory Phonetics: Theoretical Basis for LID

Articulatory phonetics analyzes the phonemes in terms of the linguistic features of these sounds, and relates these to positions and movements of the articulators. Phonemes can be classified by: (1) manner of articulation, (2) presence or absence of voicing, and (3) place of articulation.

- **Manner of articulation** describes different phones according to the ways the vocal tract restricts airflow. This divides the phonemes of English into the following broad categories: stop, fricative, liquid, glide, vowel and nasal.
- A phoneme is classified as **voiced or unvoiced** depending upon whether the vocal folds vibrate or not during its realization.
- **Place of articulation** refers to the location of the narrowest constriction in the vocal tract during pronunciation.

Different combinations of manner, voicing and place result in different phones.

The Differences Among Different Languages: the Possible Information Sources

- The phoneme sets of different languages are different.

  Languages do not use all possible combinations of manner of articulation, voicing

and place of articulation. Since each language uses only a subset of the phones, different languages tend to have different phoneme inventories.
For example, English has both voiced and unvoiced stops while Mandarin Chinese only has unvoiced stops. French has 15 vowels while Spanish has only five vowels. German has front rounded vowels and Russian has back un-rounded vowels while these kinds of phonemes are not permitted in English.

- Even when phonemes are common to two languages, they may differ slightly in realization.
  For example the German phoneme /h/ is sightly different from the English /h/: the latter is fricated less than the former.

- The lexical structures and grammars are different for different languages.
  This provides high level information for discrimination.

- Stress, duration and pitch are used differently in different languages.
  For example, Mandarin Chinese is called a tone language: different pitch patterns on the same phoneme string denote different words.

From the above sampling of articulatory phonetics, we can draw the following conclusions as a theoretical basis for LID:

- The sets of phonemes of different languages are different.

- The possible phoneme strings (legal combinations of phones, guided by the orthography and grammar of the language) are different for different languages.

- Each language has its own way to control duration, pitch and stress (prosodic information).

- Acquisition of high level knowledge about the languages can lead to near perfect performance (e.g. a native speaker can identify his or her mother language without any difficulty), since all these information sources combine to make understanding possible.

1.1.2 The Difficulties: Challenges to LID

In the previous subsection we discussed the differences between different languages. In this subsection we will discuss why language identification remains a challenge in spite of these great differences.

As described above, a spoken sentence can be viewed as the concatenation of phoneme strings according to certain rules (orthography and grammar etc.). The speech signal can be viewed as the output of the speech production process (the realization of these phonemes by the articulators) and is thus related to both its linguistic input (e.g., sentences) and extra-linguistic sources (e.g. speaker identity). One can therefore attempt to identify language directly at the acoustic level, or by studying differences at higher levels. The first approach has achieved only limited success in various attempts ([Mut93, Zis93, NUS92]), and we therefore assume that higher levels need to be studied.

To study languages above the acoustic level, we first need to recover the linguistic input from the acoustic signal. According to common usage in speech processing, we therefore need a phone recognizer (which can recognize broad categories or fine phones, or something intermediate).

This highlights the central difficulty of LID, since we are exclusively interested in (1) speaker independent (2) continuous (3) telephone speech. These three conditions complicate speech recognition severely. Since the vocal tracts of different people are different, speaker variations in the realization of phones will be substantial. Continuous speech introduces co-articulation which decreases the distinguishability of different phonemes, and makes the detection of phoneme boundaries difficult. The telephone channel has only limited bandwidth (around 3.4 kHz); thus, the high frequency information in the speech signal is lost, resulting in an increase in the error rate for phone recognition (especially for consonants).

After we obtain a phone string, we wish to use it to distinguish between languages. For this purpose, spoken utterances can be decomposed from sentences to words, then to phonemes. But for LID, even if we successfully recover the phoneme strings, we encounter a third problem: vocabulary size.

Humans have no problem in identifying a language they understand well. Similarly, there is no doubt that a human-like language identification system could achieve nearly flawless performance, if it could have a very large vocabulary which is accurated recognized and acquire knowledge of the syntactic and semantic rules for each language in the task. With current speech techniques and computer resources, development of such a system is impractical. The reasons are:

- Speech recognition system performance is still far from human levels of performance ([YWB94]).

  Current speech recognition systems work best given strong constraints on the vocabulary size. But for unrestricted LID, no such assumption can be made, since the input will be free speech. This situation makes it impossible to construct a network to recognize all the possible words (which current techniques demand); it is therefore not possible to utilize the highest levels of information (grammar, semantics etc.).

- Collecting and selecting sufficient knowledge of multiple languages is not a trivial task.

- In order to get a robust representation of this information, large amounts of training data will be required.

- Given commonly available computers, it is still a challenge to run a single language large vocabulary, continuous speech recognizer in real time. Building a LID system based on such a technique will be even more expensive, since multiple recognizers will be involved. The tradeoff between accuracy and computational simplicity has to be considered.

Thus in order to capture the language information, a modeling unit which compromises between the need for high-level information and the limitations of speech recognition algorithm is needed. This is addressed in more detail below.

1.2 Related Work

Compared with automatic speech recognition, automatic language identification has progressed slowly during past two decades. Until recently, only a small number of papers were published[Mut93]. Early approaches that exploited the acoustic feature vectors included a filter band based approach[LD74], an LPC-based polynomial classification approach[CI82], and formant vector quantization based approaches [Foi86, GMW89], House and Neuburg[HN77] 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[LE80]. With the improvement of speech processing techniques, more sophisticated approaches have been studied recently. These include systems using vector quantization [Sug91], Hidden Markov Models (HMM) [UN90, SAG91, RMM91, NUS92, Zis93], neural-networks [MC92], embedded keywords [RR94], syllabic spectral features[Li94], the distinction between polyphonemes and monophonemes [ADB94], and subword recognition exploiting acoustic/phonotactic constraints [MBA$^+$93, TCP94, HZ93, HZ94, RSN94, LG94, KH94. ZS94, BABC94]. Systematic reviews of language identification activities can be found in[Mut93, MBC94].

These efforts have improved the performance of LID systems dramatically. It was shown that accurate classification could not be achieved by simply using frame-based acoustic feature vectors; discriminates 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.

1.2.1 Early Work: 1973–1992

A detailed review of LID research during this period can be found in [Mut93]. Here, we only list a few papers which are more relevant to this dissertation.

- House and Neuberg

  By using the phonetic transcriptions of text from eight languages (*English, Chinese, Greek, Japanese, Korean, Russia, Swahili* and *Urdu*), House and Neuberg demonstrated that excellent language identification could be achieved by exploiting phonotactic information. In their paper[HN77], they showed that the phonotactic information was sufficient for language discrimination by modeling the phone sequence information as a Markov process.

  Although their experiments were carried out at a symbolic level (as opposed to on the acoustic signal), their work had great influence on recent works (e.g. [HZ94, YB95b]). Their conclusions form a basis for the current study.

- Foil

  Foil[Foi86] used formant frequencies (described in terms of values and locations) to represent the characteristic sound patterns of a language. Language identification was performed by exploiting the frequency of occurrence of these patterns. A *k-means* clustering algorithm and vector-quantization were used.

  On a three-language task, the LID correct rate was 64% with 11% rejections (on data collected from a radio receiver with a 9db SNR).

- Nakagawa, Ueda and Seino

  In their work[NUS92], four approaches (vector quantization, discrete HMM, continuous density HMM and mixture Gaussian distribution model) were studied. All the approaches were designed to perform LID by using acoustic features. The comparative experiments were conducted on the same data from four languages: English, Japanese, Mandarin Chinese and Indonesian. They found that continuous HMMs and mixture Gaussian models (correct rate 81.1%) were superior to VQ (correct rate 77.4%) and discrete HMMs (correct rate 47.6%).

  By enhancing the best system with duration modeling and dynamic features, the best result was 86.3%. However it was found thereafter that improved performance could be obtained by incorporating supra-acoustic information[MBA$^+$93, ZS94], and these purely acoustic methods have therefore become unpopular.

1.2.2 Current Activities: 1992–present

Since the release of a publically available LID database ([MCO92] etc.), research on LID has been rejuvenated, and many more papers were published during this period. Impressive results (e.g. [Zis95, Li95, YB95c]) have been achieved on fairly large tasks (relative to the available computation power).

Two systems (one developed at MIT Lincoln Lab and the other developed at ITT) achieved remarkable performance on a public evaluation (by the National Institute of Standard and Technology (*NIST*)) in 1994[MBC94]. The results they achieved are summarized in Table 1.1.

Table 1.1: NIST'94 Evaluation: The performance of systems from MIT Lincoln Lab and ITT

Task		Six-Language	Eleven-Language
MIT Lincoln Lab	45-second	90%	80%
	10-second	82%	70%
ITT	45-second	86%	78%
	10-second	75%	64%

MIT Lincoln Lab's approach[ZS94, Zis95] is called a *PRLM-P* system: multiple phone recognizers (PRs) feeding n-gram language models (LMs) running in parallel (hence the -P). The system is composed of two parts (for an N language task): (1) multiple (M) language-dependent phone recognizers (front end), and (2) M sets of N language models (see Section 1.3 for a similar system). In [ZS94], $M = 6$, and in [Zis95] $M = 16$ (gender-dependent recognizers are used for the eleven-language task). For each testing utterance, the final LID likelihood scores for each language in the task were calculated as the sum of the corresponding individual log likelihoods from each of the phone recognizers.

ITT used a speaker based system[Li94, Li95]: a nearest neighbor scoring technique is employed with score normalization processes similar to the likelihood ratio scoring at

both speaker and language levels. The input vectors are syllabic based feature vectors, containing the dynamic spectral (and prosodic) changes at syllabic nuclei. The system collects a number of feature templates for different speakers with known gender. The front end uses combined "onset" and "coda" spectral features as well as the syllabic "prosodic" features. During testing, the N nearest reference speakers (reference templates) are calculated, and the final LID result is obtained according to the majority votes.
The major differences between these two systems are:

- One (MIT Lincoln Lab) employs language statistics, while the other (ITT) contains non-parametric LID models (speaker templates).
- ITT's approach exploits acoustic information directly while MIT Lincoln Lab's approach uses information based on phone recognition.
- MIT Lincoln Lab's system is based on speaker-independent phone recognizers, and thus ignores the speaker-specific information. One of the assumptions of the ITT system, on the other hand, is that the difference between different speakers might be larger than the difference between different languages.

Note that:

- Although these two approaches use different feature sets and methods, they perform with similar accuracy.

  This suggests that with current techniques, the differences between different languages can be detected at different levels. It may be possible to further improve performance by combining several information sources.
- From the development and recent improvement of these two systems ([ZS94, Zis95] and [Li94, Li95]) and other recent reports (e.g. [KH95, HZ94, TCP94, PC95]), there is a common trend that more and more detailed modeling is employed in order to improve the existing systems. As a result, more training data are needed. Also, the computational complexity of these systems is very high. How to solve these new problems poses additional challenges to LID research.

1.2.3 The Problems

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.*[MBA$^+$93] reported significant improvement by using phonemes rather than broad categories as basis for LID, and others (e.g. [ZS94, HZ94]) have similarly found that more detailed modeling is beneficial. However, increasing the amount of details in the models inevitably requires more training data for the models to achieve the same robustness.

2. How to model prosodic information.

   Perceptually, variability in prosodic information is one of the major distinctions between different languages. How to differentiate the speaker-specific and language-specific information and how to model it effectively are still unsolved problems in speech research.

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

   The LID scores calculated from different information sources (LID models) for a given input utterance reflect different aspects of the utterance. A faulty assumption in combining these information sources may fail to capture the underlying relations.

4. How to maximize the useful information obtained 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.

5. How to decrease the computational complexity.

   For many potential LID applications, computational simplicity is important. How to balance the recognition accuracy and CPU cycle requirement is an important issue.

1.3 An Approach to Language Identification based on language-dependent phone recognition.

In this dissertation, we describe the development of an LID system based on phone recognition which addresses the above problems. Figure 1.2 illustrates the flow chart of the system.

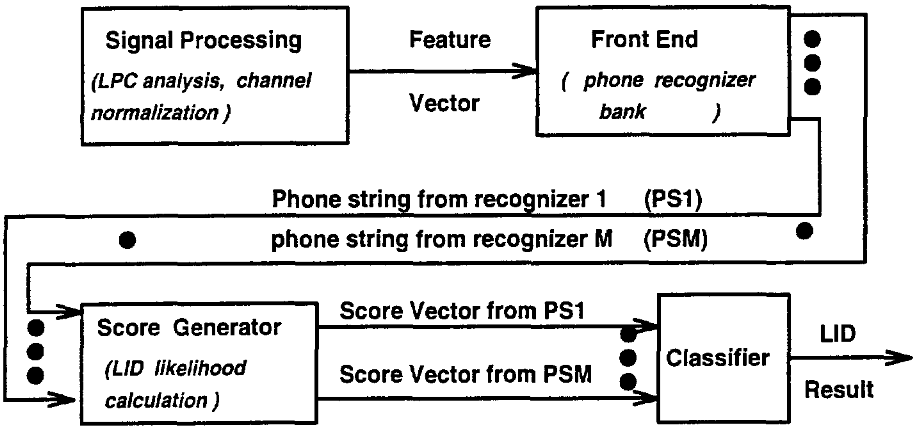

Signal Processing
(LPC analysis, channel normalization)
Feature
Vector
Front End
(phone recognizer bank)
Phone string from recognizer 1 (PS1)
phone string from recognizer M (PSM)
Score Generator
(LID likelihood calculation)
Score Vector from PS1
Score Vector from PSM
Classifier
LID
Result

Figure 1.2: Flow chart of the system

The baseline system is designed to meet the following criteria:

- A general architecture which should be able to support automatic training and testing for different tasks must be employed.
- Multiple information sources must easily be included.
- Computational complexity must be reasonable.

1.3.1 Finding a Good Modeling Unit

To satisfy these criteria, we have selected the phone as our modeling unit. It has the following advantages:

1. Natural representation of the acoustic signal.

   The phone is the realization of the linguistically fundamental unit, the phoneme.

It is therefore the bridge between the acoustic signal and high level information. From the boundaries of the phone segments and the sequences of phones, acoustic information, prosodic information and phonotactic information can be extracted.

2. Minimum representation.

   Unlike other units (such as syllables or words), the context-independent phone inventory is one of the smallest sets which can be used to describe a spoken language. Consequently the statistically based model size (number of parameters) will be small and the model itself will be trained easily. For example, for a model with N model units, the number of bigrams is $O(N^2)$. A bigram model based on words will be much larger than the model based on phones.

3. Trainability.

   Phone recognition systems are relatively easy to train. Given the effort in the past few decades, the technology for phone recognition is relatively mature (although accuracy remains an issue).

   Also, the training of monophone models does not require a very big database, and the computational complexity is relatively low both for training and testing.

4. Extensibility.

   High level information (language modeling etc.) can be retrieved from the phone string. Both prosodic information and the phonotactic constraints of different languages can be modeled in terms of phone sequences.

1.3.2 The Baseline System

The baseline system is designed based on the above considerations. The system, which is shown in figure 1.3,

is composed of three parts: (1) HMM based phone recognizers as front end, (2) LID score generators and (3) a final classifier.

Using the phone set of one language to model the phonotactic constraints for all the languages in the task has been proven to be feasible for language identification by other

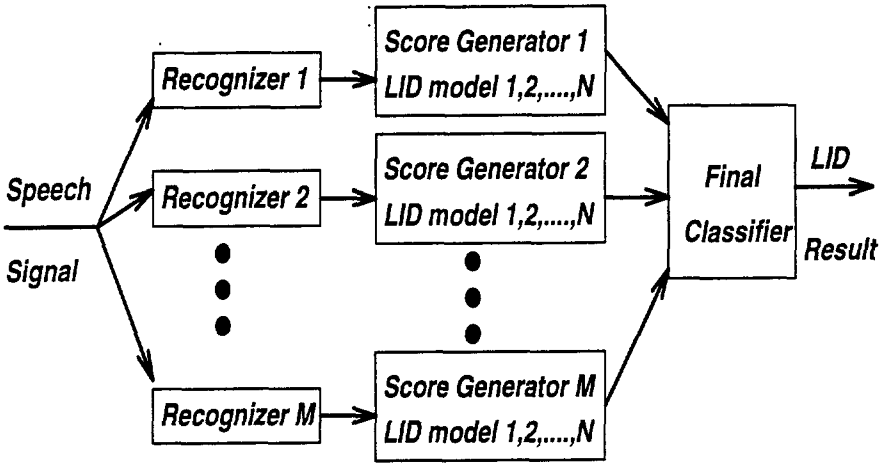


Figure 1.3: The baseline system

researchers (e.g. [MBA+93, ZS94, HZ93] etc). In this dissertation, this idea is extended.

1.3.3 Contributions: Methods Proposed to Improve the Baseline System

The major contributions of this dissertation are:

1. Increasing phonotactic modeling accuracy without drastically increasing the amount of required training data.

   A language model based on right-context bigrams is proposed as an additional set of features to the conventional forward bigram language model (which uses left contexts). The introduction of this model improves modeling accuracy (to capture both forward and backward information) without increasing the number of parameters in the model excessively.

2. A better representation of duration information.

   A straightforward extension of the duration model (a generalized context-dependent model interpolated with a context-independent model) is proposed.

3. A better way to integrate multiple information sources.

   Experiments were conducted to evaluate using a neural-network for information integration in comparison with a more standard linear classifier.

4. A new way to optimize the LID models.
   In order to compensate for limited phone accuracy, a method of optimizing the LID models based on error back-propagation is proposed.

5. Post-processing for conversational speech and new language adaptation.

At the acoustic level, cepstral mean subtraction was implemented to enhance the system robustness.

The LID systems were evaluated on six-language, nine-language and eleven-language tasks. The best results achieved were: for 45-second long utterances and 10-second long utterances, 92% and 83% (LID correct rate) for the standard six-language task, 91% and 77% on the standard nine-language task, and 90% and 77% for the standard eleven-language task. The results compare favorably with previously reported results on the same tasks.

1.4 Outline of the Dissertation

In Chapter 2, we first present comparative experiments using some previously reported approaches. and use this as foundation for the preliminary baseline system. Thereafter, the database used in this study is described (Chapter 3), and Chapter 4 describes the baseline system. Chapter 5 details the four methods proposed to enhance the system performance, and Chapter 6 presents the evaluation of our system. Our preliminary studies into the processing of conversational speech and adaptation to handle new languages are reported in Chapter 7. Concluding remarks and a description of future work are given in Chapter 8.

Chapter 2

Preliminary Study: Comparative Experiments

At the time the author began working on this thesis, a few sites [ZS94, HZ94, Li94, NUS92, MBA$^+$93] had reported respectable results. These approaches belong to two categories, using either acoustic feature-based likelihoods or language-dependent language modeling scores. These systems had been developed and evaluated using different corpora, which made comparisons between them difficult.

As an initial attempt to understand the behaviour of different approaches in these two categories using the same data set, we implemented the following approaches: (a) Gaussian mixture-based Markov Model (GMM) approach, (b) HMM-based Broad-Category (BC) classification, (c) HMM-based Fine-Phonetic (FP) classification and (d) Bigram-based Phoneme Recognition Language Modeling (PRLM) approach. These four representative approaches encompass a large part of the historical and current approaches which are related to this work.

Since the cost (in terms of CPU time and disk space) of implementing an automatic language identification system is relatively high, we limit our task to the distinction between English and Japanese utterances from a particular corpus (see below).

Based on analysis of our results using these approaches, our preliminary baseline approach which integrates acoustic level scores and language-modeling scores is evaluated. The baseline system developed using the same task data achieves our best result. This shows the importance of combining features from different sources.

2.1 Database and Feature Representation

The data used were taken from the OGI_TS corpus[MCO92]. It is a telephone speech database which was designed for language identification research. More details on this database will be introduced in Chapter 3. We used the 45-second long "story-before-the-tone" (story-bt) parts in English and Japanese which had been phonetically labeled. For each language, the training set contained 80 utterances, the development-test set contained 20 utterances, and the test set contained 50 utterances. The ten-second utterances were obtained by chopping the whole utterances. Since some story-bt utterances were shorter than 45 seconds, we had 382 utterances total in our 10-second test set.

Speech data were parameterized every 25.6 ms with 12.8 ms overlap between contiguous frames. Two independent streams of feature vectors with 13 dimensions each were calculated: 12th-order Mel-Scale LPC Cepstra appended with normalized Energy and the Delta Cepstra appended with Delta Energy.

During identification, the log probability at the acoustic level for language l is calculated as:

$$P(\{c_t, d_t\}|M_{cl}, M_{dl}) = \sum_{t=1}^{T}(C_1 logP(c_t|M_{cl}) + C_2 logP(d_t|M_{dl})) \qquad (2.1)$$

where c_t and d_t are the *cepstra + energy vector* and the *delta vector* respectively, M_{cl} and M_{dl} are the statistical models for the corresponding vector streams for language l in the task, and C_1 and C_2 are the stream weight coefficients. In our implementation, we set $C_1 = 1$ and $C_2 = 0.6$.

2.2 Comparative Experiments

In this section, the algorithms compared are detailed.

2.2.1 Gaussian Mixture based Markov Model Approach (GMM)

In GMM approaches[NUS92, Zis93], the sequence of frame-based feature observations extracted from the speech signal is modeled by a Markov process. The Markov process is formalized mathematically as a set of state transitions within a given state space. The transition probabilities depend on the current state and the observation, and are independent of previous observations and transitions. Thus, if the observation vector at the time frame is represented by X_i and P_i is the probability of the vector being in the various states at time i, we can write:

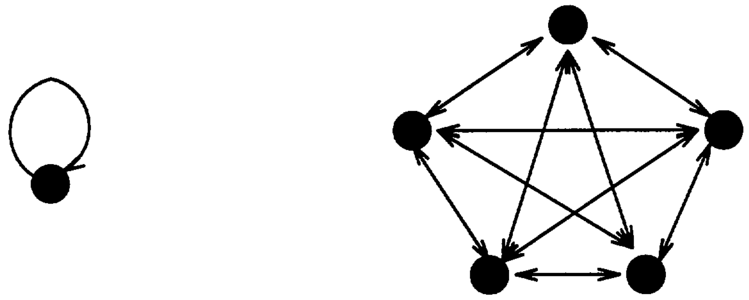
$$P_{i|1\ldots i-1}(X_i|X_1, \ldots, X_{i-1}) = P_{i|i-1}(X_i|X_{i-1}) \tag{2.2}$$

The Markov process has been used extensively in speech processing, especially in noise reduction, speaker identification, accent detection and language identification.

The advantage of this approach is its simplicity, both in training and testing. Importantly, it does not need labeled data. It allows us to find the distinction between different languages based on the acoustic observations alone.

Each language is modeled by a Markov Model, with Gaussian Mixtures associated with each stream on each state. During testing, a maximum-likelihood algorithm is applied for each language model given the test data in order to find the maximum likelihood path. The language is identified based on a comparison of the maximum likelihood values obtained from the models of each language.

Two experiments were carried out, using the Markov topologies given in Figure 2.1.



(a) One state

(b) Five state

Figure 2.1: Two Markov Model topologies for GMM approach

The general architecture of the system is given in Figure 2.2.

- Single state per language

  In this experiment, the state space of each language contains only one state. A 36-element Gaussian mixture is associated with each stream.

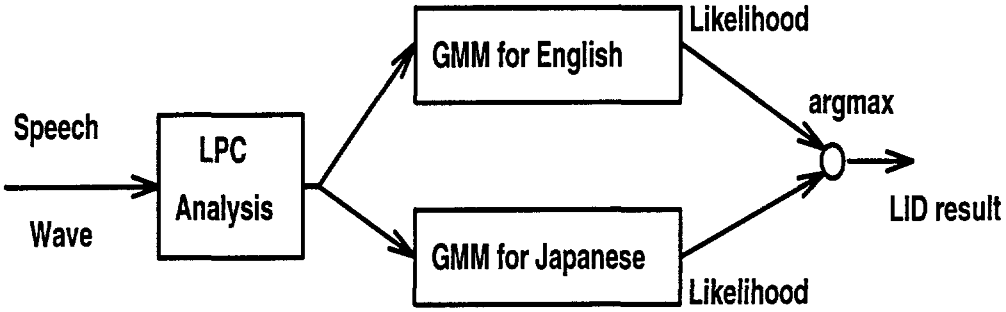


Figure 2.2: System Architecture for GMM approach

- Five states per language

  In this experiment, the state space of each language contains five states. The transition probabilities between all states are initially equal, and the Gaussian mixture of each stream for each state consists of 36 tied global Gaussian density functions.

During training, the training data are randomly chopped to a length of between 0.1 to 3 seconds length. A forward-backward algorithm was run on these data until the models converged to the preset threshold.

The Viterbi algorithm is used to decode the test data. The final decision is made based upon the maximum likelihood criterion given in (2.3):

$$l = argmax \sum_{t=1}^{T} P(\{c_t, d_t\}|M_{cl}, M_{dl}) \qquad (2.3)$$

2.2.2 Broad-category based Approach (BC)

A broad-category based approach to automatic language identification has been studied by Muthusamy[Mut93]. In our approach, the phonemes of each language are divided into six broad categories, *i.e., vowel, consonant, nasal, liquid, noise and non-linguistic.*

The *non-linguistic* model is proposed in this work since phenomena such as the filled pause are very common in natural (as opposed to read) speech. Within the language, these phonemes are acoustically different from their phonemic counterparts in normal continuous utterances. Furthermore, different classes of filled pauses are typically used in different spoken languages. We thus decided to model filled pauses explicitly to increase the robustness of the language model (since they are the "phonetic outliers" within the phonemic classes). An informal test showed this model increased system performance slightly. In natural speech, silence segments are also sometimes too long to be absorbed by the phoneme models, and noise levels are not consistent due to the variable communication channels. We therefore also group noise and extended silence segments together as a single *noise* model.

We did not divide the consonant class into the traditional stop subclass and fricative subclass since we consider the loss of consonant information due to the bandwidth of telephone channels to be relatively large. Hence distinctions between these subclasses tend to be unreliable and less informative.

One broad-category phone (class) recognizer is trained for each language. The transition probabilities between pairs of classes are estimated from the training data and smoothed using the development data. The smoothing function we used is given in (2.4):

$$P(S_j|S_i) = \alpha P_t(S_j|S_i) + (1-\alpha)P_d(S_j|S_i) \tag{2.4}$$

where S_i, S_j denote different classes, P_t and P_d denote the log transition probabilities from class S_i to class S_j which are calculated from training data and development data respectively. α is the smoothing factor; in our approach it is set to 0.8, which represents the ratio of the sizes of the two data sets.

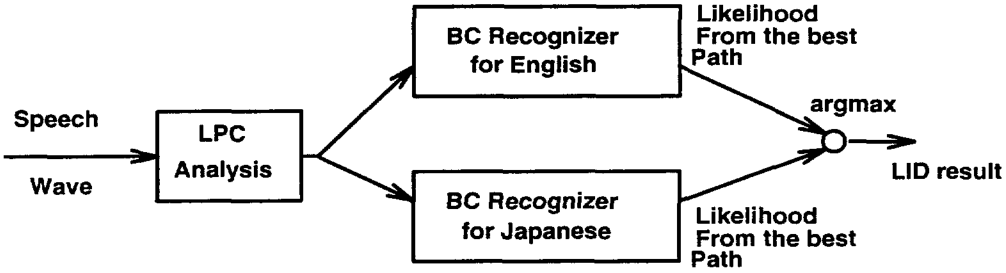


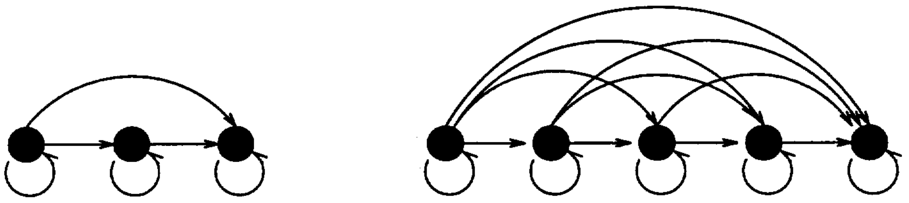
Figure 2.3: System architecture for BC approach

The HMM model for each broad-category phone (class) for each language has a four-element Gaussian mixture associated with each stream on each state. A Gaussian density function is represented by a mean vector and full covariance matrix. Language identification is performed using a beam searching Viterbi algorithm decoding the test utterances once by the broad-category recognizer of each language. The two *log* likelihood scores obtained from the maximum likelihood path decoded by the Viterbi search for each language are used in the maximum likelihood criterion stated in (2.3) to make the final classification.

The system architecture is given is Figure 2.3.

The pruning thresholds used in the Viterbi beam search were adjusted independently for each language based on the classification performance on the development set. This type of adjustment is performed throughout the following approaches, and not mentioned explicitly.

The accuracies of the broad-category class recognizers for English and Japanese on the development-test set were 59.3% and 60% respectively.



(a) Three-state HMM

(b) Five-state HMM

Figure 2.4: Two HMM topologies for BC approach

Two HMM topologies for each broad-category phone (class), which are shown in Figure 2.4 , are trained and tested.

2.2.3 Fine-phonetic Approach (FP)

As with the BC approach, we process an utterance of unknown language in parallel by different sets of models (language-dependent phone recognizer) for each of the languages in the task, and choose the language associated with the recognizer providing the highest likelihood score. Here, however, phone models rather than broad-category models are used ([MBA$^+$93, LG93, ZS94, RR94]).

In our implementation, the system configuration (shown in Figure 2.5) is similar to the configuration of the broad-category based approach. The differences between these two approaches are:

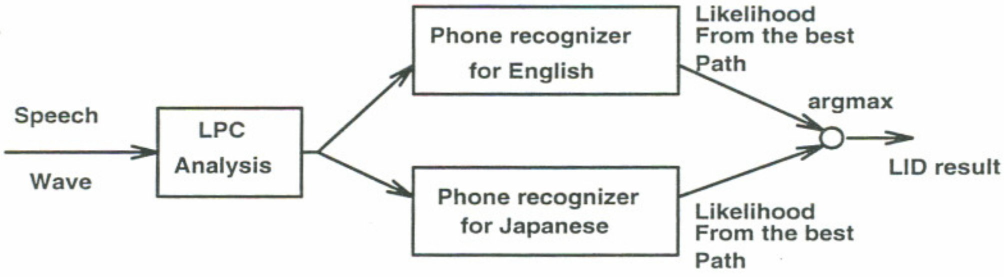


Figure 2.5: System Configuration for FP Approach.

1. For each language, a fine phone recognizer is used instead of a broad-category class recognizer. Each phone recognizer of a language contains the models for its mono-phones (some mono-phones are merged because they are rare in the training data), and the most frequent right-context-dependent biphones (those which occur more than 70 times in the training data). This definition results in 175 models (46 mono-phones and 129 right-context-dependent biphones) for English and 119 models (26 mono-phones and 93 right-context-dependent biphones) for Japanese.

2. In each state of any phone model, a three-element Gaussian Mixture is associated with each stream; the Gaussian density function is represented by a mean vector and diagonal covariance matrix.

3. The bigram grammars used with the Viterbi search algorithm are estimated from the training data and smoothed with the development data.

4. After the Viterbi search finds the most likely path, the probabilities and frame numbers associated with each segment except the noise are accumulated; the final likelihood scores for each language are obtained by normalizing the accumulated probability with the accumulated frame number.

The phone accuracies for the English and Japanese phone recognizers were 48% and 58% respectively.

2.2.4 Bigram-based Phoneme Mapping Approach (PRLM)

Here we use a language-dependent phone inventory to model the phonotactic constraints of all the languages in the task (see Figure 2.6). Thus, several language-dependent phone recognizers run in parallel([ZS94]) to enhance system performance.

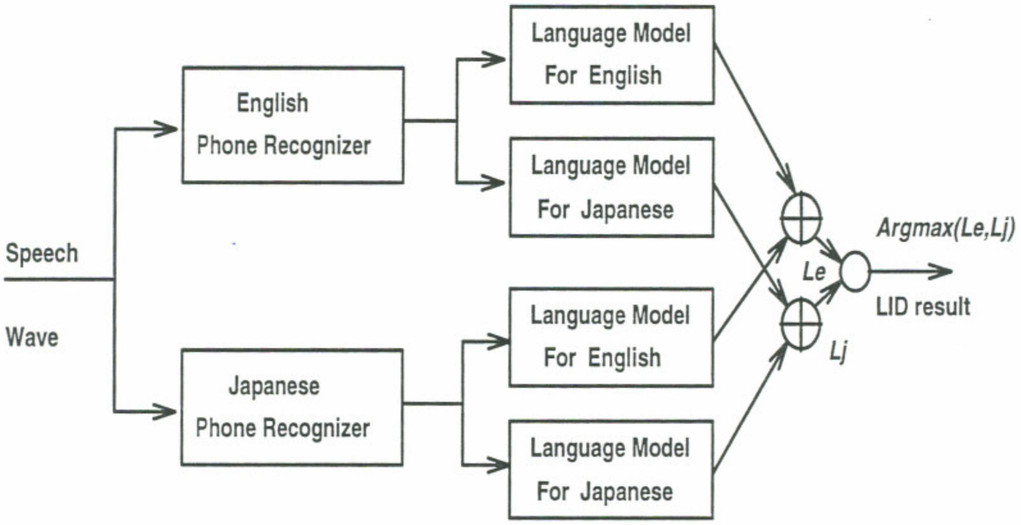


Figure 2.6: System Configuration for PRLM Approach.

In our implementation, the similarities between this approach and the previous approach are:

1. Each language has its own phone recognizer.
2. The topology of each phone model is the same.
3. The same Viterbi algorithm is used to find the most likely path.

The differences between these two approaches are:

1. Unlike in the FP approach, only mono-phone models are used in this approach.
2. A language model based on bigrams and unigrams[Jel90] is used to capture the phonotactic constraints:

$$P = \sum_{i=1}^{T}(\alpha P(O_i|O_{i-1}) + \beta P(O_i)) \tag{2.5}$$

where O_i is the ith phone in the decoded best path, $P(O_i|O_{i-1})$ is the bigram term and $P(O_i)$ is the unigram term. α and β are the weight coefficients; in our approach, they are empirically set to 1.0 and 0.6 for the whole utterances and 1.0 and 0.2 for the ten-second segments respectively, and log probabilities are used. The bigram terms are estimated from the training data and smoothed using the development data, and the unigram terms are calculated from the relative frequencies of the phones in the (decoded) training data.

3. Associated with each phone recognizer, one language model is generated for each language through phone mapping. The models are obtained from the training data decoded by the associated recognizer. During testing, each recognizer generates a score for each language. The scores are accumulated for each language from the scores generated by each recognizer.

4. Unlike the previous approaches, the quantitative measurement obtained at the level of acoustic features is discarded. Only the scores obtained from the language models contribute to the final decisions.

2.3 Preliminary Baseline Approach

Our preliminary baseline approach is introduced in this section. It incorporates several sources of information within one system framework. In particular, acoustic and language modeling features are combined.

2.3.1 Motivation

As we discussed in the previous section, language modeling is a powerful technique in LID, but system performance depends solely on the performance of the phone recognizer, *i.e.*, depends on the accuracy of the decoded phone strings. The segment-based acoustic features (*e.g. cepstrum*) focus on the realization of each individual phone. From the error analyses of the comparative experiments (which we will discuss in section 2.4), we know that these two information sources are not entirely correlated.

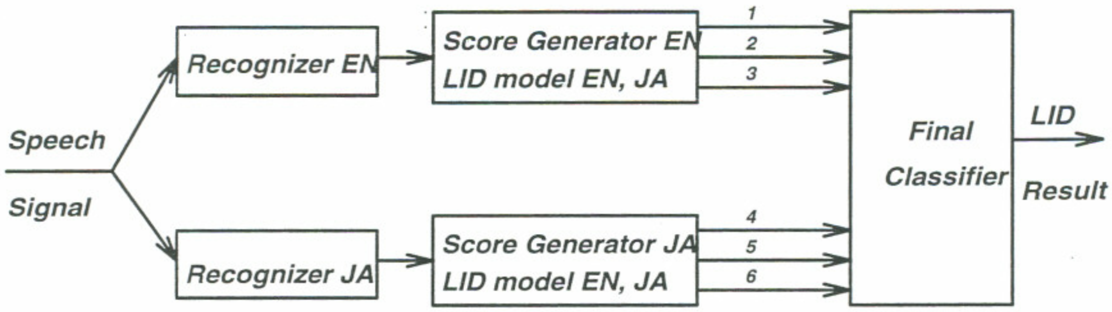


Figure 2.7: Our baseline system for these two language task.

2.3.2 The Baseline System Structure

The baseline system for this task is shown in Figure 2.7. The system consists of *three* parts: phone recognizers for English and Japanese, two sets of language models for these two languages for each corresponding recognizer, and a final classifier.

The arrows labeled *1,...3* represent the acoustic and phonotactic scores for being English and Japanese based on the output of the English front end. Similarly, *4* to *6* represent the same scores based on the output of the Japanese front end.

2.3.3 Implementation

1. **Front end**

   The front end has two functions: phone recognition and segmentation. For an N-language LID system we implement N phone recognizers to capture both acoustic and phonotactic information for each language. Considering implementation efficiency and the compromise between these two functions, we select an HMM as our front end. The HMM is a good technique for continuous speech recognition though it is not an accurate boundary detector (segmentor).

   The HMM-based phone recognizers for this experiment were those used in the PRLM-P experiment.

2. **Score generator**

   The implementation of the score generator for each front end is straightforward.

During training, we estimate the language models (n-gram) as given in (2.5) and the acoustic model of each phone for each language from the training data. During testing, test scores are calculated from the output of the front end.

For this two-language task, we have two score generators. Each score generator will generate 2+1 scores (two language modeling scores, one acoustic score). There will thus be *six* inputs to the classifier.

The language model contains only monophones; before we estimate the language modeling score, the decoded context-dependent phonemes are converted back to monophones.

3. **Classifier**

   A linear classifier[Fuk90] was used. The input to the classifier for an N-language LID system will be two vectors (N^2 language modeling scores and N acoustic likelihood scores), and the physical meanings of these two vectors are different. A classifier should have the ability to learn the relations among these scores.

2.4 Results and Analyses

All the results of the approaches described in this chapter are given in Table 2.1.

We see that system performance is highly correlated with system complexity. A detailed examination of the misclassified utterances revealed that for the first three approaches (which rely upon various forms of the acoustic likelihoods), the sets of misclassified utterances overlapped substantially, which confirmed the need for detailed modeling.

The GMM results using one-state and five-state models are similar. The number of Gaussian mixtures is clearly not large enough to capture the information of all phonetic realizations; therefore the temporal information in the signal can not be exploited, resulting in a failure to capture the sequential information by the multi-state approach.

From the results of the two broad-category experiments, we found that with limited training data, a three-state phone model is sufficient. Therefore the three-state model is used in PRLM and our baseline approaches.

Table 2.1: Identification rate of all the approaches

Approach	whole utterance	ten second utterance
GMM (1 state per language)	76%	72.0%
GMM (5 state per language)	77%	72.5%
BC (3 state per class)	88%	84.0%
BC (5 state per class)	89%	84.0%
FP	90%	86.1%
PRLM	93%	87.0%
Baseline	95%	91.6%

By comparing the misclassified sets of the fine phonetic approach and the phoneme-mapping language-modeling approach, we found that about 50% of the errors did not overlap. Intuitively, the acoustic scores are more sensitive to the accents of the speakers and the histogram of the relative occurrences of the phonemes in the training data, while the language modeling scores are more sensitive to the content of the speech, *i.e.*, the phone sequences. These two information sources are thus somewhat orthogonal; combining them explicitly introduces additional information. Our baseline approach causes the error rate of the language modeling approach to decrease by 30%, which confirms our observation.

Chapter 3

Database and Task

3.1 Database

The major database used in this study is the OGI_TS database ([MCO92]), which is a multi-language telephone speech database designed for language identification research. The data were collected using a Gradient Technology Desklab recording equipment via a SCSI Port. Speakers could reach the speech equipment via a toll free telephone number. In the original database, the speech signal was sampled at 8 kHz with 14 bits resolution and the data were collected via analog telephone lines. The latest data in the database were collected via digital lines and were 8 bits mulaw encoded. A fixed recording gain was used. The recording process was controlled by questions/prompts. 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 transcribed phonetically.

This database is publically available and is used as a standard database by the National Institute of Standard and Technology (NIST) for language identification progress evaluation.

The other database used for conversational speech processing is the data sample distribution from the Call Friend Project (collected by the Linguistic Data Consortium (LDC)). These data were recoded on switchboards. The two channels (local and remote) were recorded simultaneously and stored in two different files. It was released in May 1995, with a total of 6 hours of speech data. It contains data from 9 languages, which are real telephone conversations.

3.2 Task

Since the training of our phone recognizers needs phonetically transcribed data, the maximum number of recognizers available within this database is six ($M = 6$). We evaluate the development of our approach with two tasks, namely on only those languages which are transcribed ($N = 6$), and on all languages ($N = 11$). In addition to these two tasks, the final system is also evaluated on a nine-language task ($N = 9$), for which additional test data became available recently.

In order to make the results comparable to those of other sites, we use the NIST (94 and 95) evaluation test data as our final test set.

3.2.1 Six-language Task

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

Two parts of the utterances in the database were used: "story-before-the-tone" (story-bt) and "story-after-the-tone " (story-at). Totally there are 777 utterances for training, which are divided into the following three sets.

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

In each set, the numbers of utterances from each language are approximately balanced.

The test set is the same test set used by the National Institute of Standard and Technology in their March 1994 evaluation for these six languages. It contains 112 whole utterances (nominally of 45 seconds duration each) and 370 ten-second utterances, which are also part of the OGI_TS database. None of the above four sets overlaps.

The distributions (numbers of utterances from each language) of these sets are summarized in Table 3.1 and 3.2.

Table 3.1: Summary of the six-language task training sets and development set: number of utterances

Language	English	German	Hindi	Japanese	Mandarin	Spanish
Training Set 1	50	50	50	50	50	50
Training Set 2	50	50	50	80	76	94
Development Set	10	11	18	9	15	14

Table 3.2: Summary of the six-language task test set: number of utterances

Language	English	German	Hindi	Japanese	Mandarin	Spanish
45-second	19	20	20	19	17	17
10-second	69	65	65	61	52	58

Training set 1 is used to train the language-dependent phone recognizers for these six languages. Training set 2 is used to train the LID models. The development set is used to evaluate the performance of the recognizers and all these sets are used to train the final classifier.

3.2.2 Eleven-language Task

For the eleven-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). There are totally 1785 utterances in the training data (data used in the six-language task are also included), which are further divided into the following three sets.

- Training set. It has 716 utterances.
- Development set 1. It has 651 utterances.

- Development set 2. It has 371 utterances.

Again, the numbers of utterances from each language are approximately balanced.

The test set for the eleven-language task is the test set used by NIST in the March 1994 evaluation. It contains 195 whole utterances and 625 ten-second utterances. The distribution of the data in this test set is summarized in Table 3.3 and 3.4. In the table, the name of each language is represented by its first two characters.

Table 3.3: Summary of the Eleven-language task training set and developments: number of utterances

Language	EN	FA	FR	GE	HI	JA	KO	MA	SP	TA	VI
Train Set	50	50	50	50	50	108	50	104	104	50	50
Development Set 1	60	58	62	61	67	65	52	55	69	38	64
Development Set 2	27	34	33	31	56	33	31	33	29	32	32

Table 3.4: Summary of the Eleven-language task test set: number of utterances

Language	EN	FA	FR	GE	HI	JA	KO	MA	SP	TA	VI
45-second	19	19	18	20	20	19	17	17	17	14	15
10-second	69	58	62	65	65	61	45	52	58	43	47

None of the above four sets overlapped.

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

3.2.3 Nine-language Task

The performance of a system should be measured in different conditions in order to get a comprehensive evaluation. The NIST'95 test set is used to evaluate our final system. The test set contains data from 11 languages. In the set, data from two languages are treated as background utterances. The nine languages in the task are: English, French, German, Hindi, Japanese, Mandarin, Spanish, Tamil and Vietnamese. The two background languages are: Czech and Portuguese.
We use the following tasks to evaluate our system:

1. six-language closed-set test.

   The six languages are: English, German, Hindi, Japanese, Mandarin and Spanish. "Closed-set" means the test set only contains the utterances from the languages in the task.

2. six-language open-set test.

   "Open set" means the test set contains some utterances from background languages which are not in the task. In this case, rejection should be made. In this task, the background languages are: Czech, French, Portuguese, Tamil and Vietnamese.

3. nine-language closed-set test.

4. nine-language open-set test. The background languages used in this task are Czech and Portuguese.

5. EN-L pair test.

   This is a pairwise classification experiment. The goal of this experiment is to measure the separability of language pairs by the system. As a convention ([Mut93, ZS94]), we selected English as our anchor language. The pairwise experiment measures the separability between English and the other ten languages in the test set.

The training data of this task are from the training data for these nine languages in the eleven-language task. The test set used is the NIST'95 evaluation data. There are 220 whole utterances and 801 ten-second utterances in the test set. We use this set as the final complete test set to evaluate the system on these commonly used tasks. The data set is summarized in Table 3.5.

Table 3.5: Summary of the NIST'95 test set: number of utterances

Language	EN	FR	GE	HI	JA	MA	SP	TA	VI	Other
45-second	20	20	20	20	20	20	20	20	20	40
10-second	78	77	73	71	74	73	70	71	66	148

Note that: the training and testing data for the six- and eleven-language tasks were collected from analog telephone lines and were encoded in NIST_1A short-pack format. The test data for this nine-language task were collected from digital telephone lines (T-1) and were 8 bits mulaw encoded.

3.2.4 Conversational Speech

We have to date only received a prerelease of 72 calls from the LDC call-friend database, and this was used in a preliminary study on language identification with conversational speech. The sample release data were collected in the USA and Canada. Each utterance was truncated into a five-minute long segment. The prerelease contains data from English, Farsi, German. Hindi, Japanese, Mandarin, Spanish, Tamil and Vietnamese. The sample distribution is summarized in Table 3.6.

Table 3.6: Summary of LDC sample data: number of utterances

Language	EN	FA	GE	HI	JA	MA	SP	TA	VI
utterance	10	8	2	10	10	10	10	10	2

Chapter 4

Baseline system and the Experiments

The baseline system is proposed based on the results of our comparative study discussed in Chapter 2. The implemented baseline system provides the benchmark for the studies presented in this dissertation.

4.1 Baseline LID System

In this section, we 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 system provides both a benchmark measurement and a platform for the further development of this approach.

4.1.1 Models Used in the Baseline System

The statistical models in the score generator are the foundation for system performance. Two sets of models, designed to exploit sequential and duration information in different languages, are used in the baseline system.

Although acoustic models have been proven to be useful in our preliminary study, the training of acoustic phone models (HMM) needs phonetically labeled data. For a large task (such as an eleven-language task), these transcribed data may not be available. Therefore acoustic models are not used in our system.

Language Model

House and Neuberg ([HN77]) proposed that sequential constraints on phonemes could be exploited as an efficient approach to language identification. Their work showed that the phone sequential constraints of different languages could be powerful features to distinguish different languages even when the speech events were described by broad-category classes. This idea has been used extensively in recent research ([MBA+93, HZ93, ZS94, KH94]), and reflects the phonotactic differences between different languages ([Lad93]). Previous work (e.g.[MBA+93, ZS94, HZ93]) showed that fine-phonetic categories could lead to better performance than broad categories. Also, our comparative experiment results support this conclusion. Therefore, fine-phonetic categories are used in this work.

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

$$P_{LF} = \sum_{i=1}^{T} log(\alpha P(O_i|O_{i-1}) + \beta P(O_i)) \tag{4.1}$$

where O_i is the ith phone in the decoded best path, $P(O_i)$ is the unigram term and 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 [Jel90].

Duration Model

A duration model has been used in ([HZ93, HZ94]) to capture a certain class of prosodic information in different languages. In our studies, two representations of the 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 models are estimated based on decoded phone strings in the training set.

4.1.2 System Architecture

Our baseline LID system is composed of three parts: (1) a language-dependent phone-recognizer based front end, (2) LID score generators, and (3) a language classifier. The general architecture for an N-language task is given in Figure 4.1, where M is the number of recognizers used ($N \geq M$).

Perceptual experiments on a 10-language task ([MJC94]) found that the number of languages known was a significant factor in the performance of human subjects. Similarly, several language-dependent recognizers (with bigram models for target languages) enhance the performance of our automatic system.

Similar system architectures were used in [MBA$^+$93, ZS94, KH94, MBMB95].

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 in parallel, and independently decode the input speech vectors into phone strings. This system configuration was first proposed in [ZS94].

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

   We found the performance of the LID system to increase for all the languages in the task with an increasing number of recognizers in the system, which is consistent with the conclusion of the perceptual experiment. The major drawback of this type of implementation is that computational complexity is also increased with an increasing number of recognizers.

2. LID Score Generator

   As shown in Figure 4.1, the phone recognizer of each language has its own score

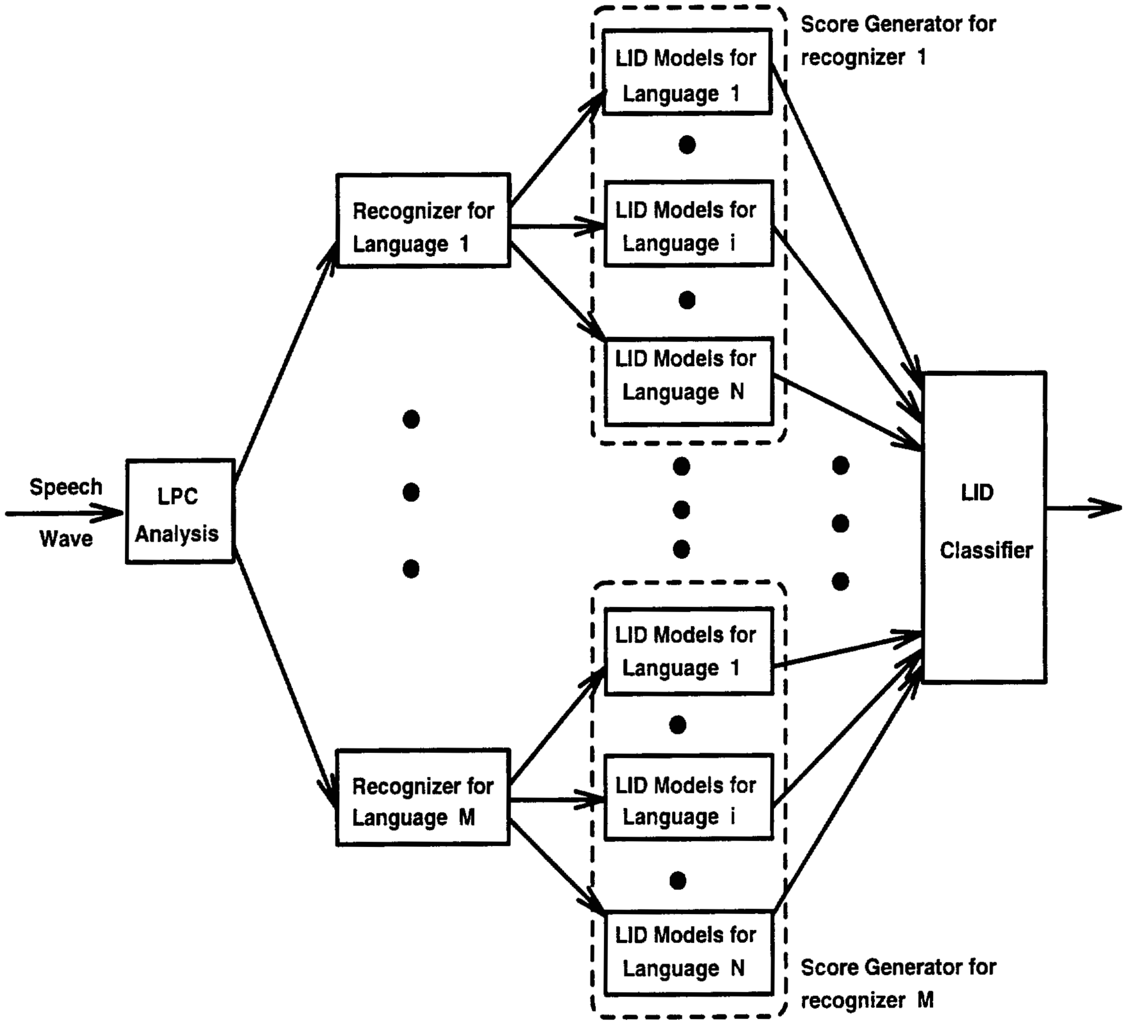


Figure 4.1: General Structure of the LID system For an N-language task with M recognizers

generator in our system. Each score generator contains a set of LID models for each language in the task. The score generator takes the outputs (the decoded phone strings) from each language-dependent phone recognizer, calculates various LID feature scores, and provides them to the final classifier.

The phone recognizers were trained on the phonetically labeled data. Decoded by these recognizers, data from the training set were used to estimate the language models and the duration models. By using each recognizer to decode the speech data from all the languages in the training sets, 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 calculates 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 language modeling scores and duration scores are normalized by the number of phones in the best path decoded by their corresponding phone recognizers.

3. Final classifier

   When multiple information sources are used, an important issue is how these 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 language scores with optimal weights.

4.2 Baseline System Evaluation

Our baseline system was evaluated on the six- and eleven-language tasks using the database described before. These results provide a benchmark for the evaluation of our further research.

4.2.1 Implementation of the Phone Recognizers

Six continuous HMM-based phone recognizers were implemented using the phonetically labeled data. Each phone is modeled as a three-state left-to-right HMM model. Three Gaussian mixtures are used to model the feature probability density function for each state in the model. Speech data are parameterized every 20 ms with 10 ms overlap between contiguous frames. For each frame a 26-dimensional feature vector is calculated: 12th-order LPC cepstra, 12th-order delta cepstra, normalized energy and delta energy. The delta feature is calculated as (4.2):

For *i* from *1* to *12* (the analysis order),

$$dCep(i) = \alpha \times \sum_{j=1}^{2} j \times (Cep(i+j) - Cep(i-j)) \tag{4.2}$$

where the dCep denote the delta features, Cep denotes the cepstra, and α ($= 0.2$) is used to scale these features.

Table 4.1: Summary of each Phone Recognizer.

Recognizer	English	German	Hindi	Japanese	Mandarin	Spanish
Size	40	37	39	27	38	28
Accuracy	46.8%	46.6%	48.1%	56.3%	36.3%	54.6%

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 4.1 gives the number of phones used in each language and the corresponding phone recognition accuracy. The phone models for each recognizer are summarized in Table 4.2.

4.2.2 Evaluation of Systems with One Phone Recognizer

One benefit of using multiple language-dependent phone recognizers is to provide a better phone coverage. In order to understand the relative importance of each phone recognizer, we evaluated the system performance by using only one phone recognizer at a time. The

Table 4.2: The phone sets used in the six phone recognizers.

FRONT END	PHONE SET
English	aa ae ah aor aw ay b ch cl d dh dx eh er ey f g hh ih iy jh k l m n ow oy p r s sh sil t th uh uw v w y z
German	aa ae ah aw ay b cl cx d ea eh eyw f g h ia ih iy k kx l m n oa oy p rr s sh sil t ts uh uu v y z
Hindi	aa ae ah ao ay b ch cl d dd dt dth eh ey f g h ih iy jh k kh l m n ng ow p r rd s sh sil t uh uw w y z
Japanese	aa b ch cl d dz ey f g h iy jh k m n ow p q r s sh sil t ts uw w y
Mandarin	ae ao aw ax ay c ch cl eh ey f h ix iy iyw k kh l m n ng oe ow p ph r s sh shr sil t th ts tsh tsr uw w y
Spanish	aa b ch cl d ey f g h iy k l ly m n ng ny ow p q rr s sil t uw w y z

evaluation was conducted on the eleven-language task.

To simplify the evaluation, after the scores for being each language in the task is calculated, the final identification result is obtained by using (4.3), where S_i denotes the score for being language i in the task.

$$l = \arg\max(S_i) \tag{4.3}$$

The identification rates (correct rates) for each language in the task by each language-dependent phone recognizers are summarized in Figure 4.2.

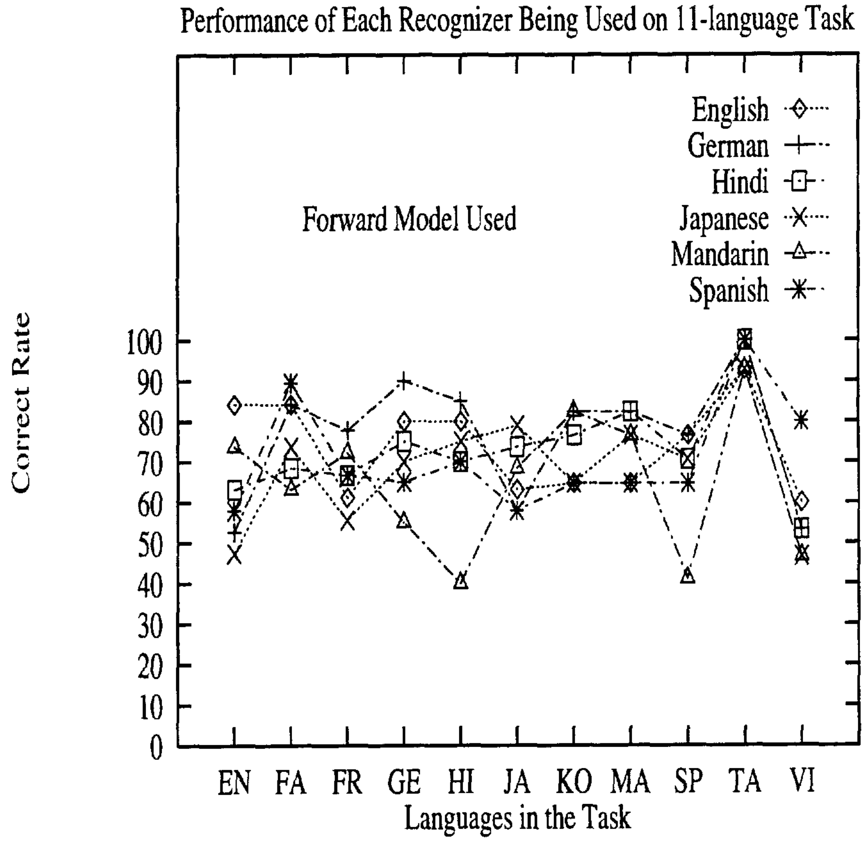


Figure 4.2: LID performance by each recognizer based on the forward language model

It shows that the correct rates for different languages by systems with different phone recognizers are different. For example, the system implemented with the English phone recognizer is relatively good at identifying English while it is not good at identifying Vietnamese; the system implemented with the Spanish recognizer is relatively good at identifying Vietnamese while it is not good at identifying English.

The overall identification rates using each phone recognizer, and also the performance when all recognizers are combined, are given in Table 4.3.

Table 4.3: Overall performance of systems based on one phone recognizer: using forward language model

Recognizer	English	German	Hindi	Japanese	Mandarin	Spanish	ALL_SIX
Mean	73.8%	76.4%	72.3%	68.2%	64.1%	70.3%	81.6%
S.D.	10.8%	14.7%	11.2%	13.3%	16.5%	12.7%	9.5%

"S.D." in the table refers to the standard deviation of the identification rates among different languages. Note the substantial improvement obtained when all recognizers are combined.

4.2.3 Benchmark Evaluation

The language models and the duration models used in the baseline system were trained from the decoded training data (phone strings) produced by the six phone recognizers. 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 4.4.

The relative importance (in terms of system performance) of each model set in the score generator is measured in the first two experiments. The third experiment gives the benchmark performance of our baseline system when both LID models are used.

Table 4.4: Baseline System Evaluation: The Benchmark Results

Task		F	D	BS
Six Language	45-second	84.8%	55.4%	86.6%
	10-second	74.1%	45.1%	76.2%
Eleven Language	45-second	76.4%	42.1%	78.5%
	10-second	65.1%	32.0%	67.0%

Compared with the results listed in Table 1.1, which were published in 1994, the performance of our baseline system is still mediocre. In the next chapter we will discuss our efforts to improve the system performance.

Chapter 5

Methods Used to Improve the Baseline System Performance

In this chapter, we present the four methods we have developed to improve the performance of our baseline LID system. These methods are designed to cope with the five important issues for LID approaches based on phone recognition discussed in Chapter 1. Also, we report on the use of channel normalization to improve robustness.

5.1 Enhanced Language Model

Language modeling is the key part of approaches which exploit phonotactic constraints. In this dissertation, a novel language model is proposed.

5.1.1 Forward and Backward Bigram-based Language Model

The language model used in our baseline system to exploit left-context information is based on the interpolated N-gram model[Jel90] as shown in (4.1); 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 a trigram-based language model for a single language in a recent effort[KH94]. 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 model 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. The backward bigram language model used in this thesis work is given in (5.1).

$$P_{LB} = \prod_{i=1}^{T}(\alpha P(O_i|O_{i+1}) + \beta P(O_i)) \tag{5.1}$$

One possible way to combine these two bigram models into a language model is as:

$$P_{LFB} = \prod_{i=1}^{T}(\alpha P(O_i|O_{i-1}) + \beta P(O_i|O_{i+1}) + \gamma P(O_i)) \tag{5.2}$$

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 are thus traded off in a way that improves our modeling of the phonotactic constraints.

5.1.2 Comparison of our Model with the Trigram Model

The commonly used trigram-based interpolated language model is given in (5.3).

$$P_{LT} = \prod_{i=1}^{T}(\alpha P(O_i|O_{i-1}, O_{i-2}) + \beta P(O_i|O_{i-1}) + \gamma P(O_i)) \tag{5.3}$$

Here we give an example to show how the training data requirement is decreased if (5.2) rather than (5.3) is employed.

Table 5.1: Comparison of OGI Model and Trigram Model

MODEL	NO. OF PARAMETERS	TRAINING TOKENS/PARAMETER
Trigram	40 x 40 x 40 = 64000	0.35
OGI Model	40 x 40 x 2 = 3200	7.03

For simplicity, only terms with the highest order in (5.2) and (5.3) are considered. For a language with 40 phones, suppose we have:

- 50 45-second long training utterances.
- each utterance has 450 phones.

So there are 50 x 450 training tokens total.

The comparison between our model and a trigram-based language model is illustrated in Table 5.1. We see that , given a limited amount of data, (5.2) allows for a relatively robust estimation of the model parameters for the proposed language model.

5.2 Enhanced Duration Model

It has been argued that duration information is useful in language identification (e.g. [HZ93, HTG95]). On the other hand, speech rate is a speaker-dependent influence which renders duration information less language specific. Based on an analysis of the data, a generalized context-dependent duration model is proposed here, to extract as much language-dependent information as possible from phoneme durations. It is a natural extension for the traditional context-independent model.

5.2.1 Analysis of the Duration Information

In order to understand the potential differences between different languages from duration information, we calculated the mean and variance of each phone (as decoded by the English phone recognizer) from different data sets. The means are summarized in Figure 5.1 and Figure 5.2. The phone index in the figures is given in Table 5.2.

We see that, for phone duration, the inter-language differences in phone duration are smaller than the intra-language differences. Although the variance in the duration of each phone is substantial, duration is still an interesting feature that can be exploited for language identification. Hence, we think a more complicated duration model (compared with the one used in our baseline approach) is desired.

5.2.2 A New 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

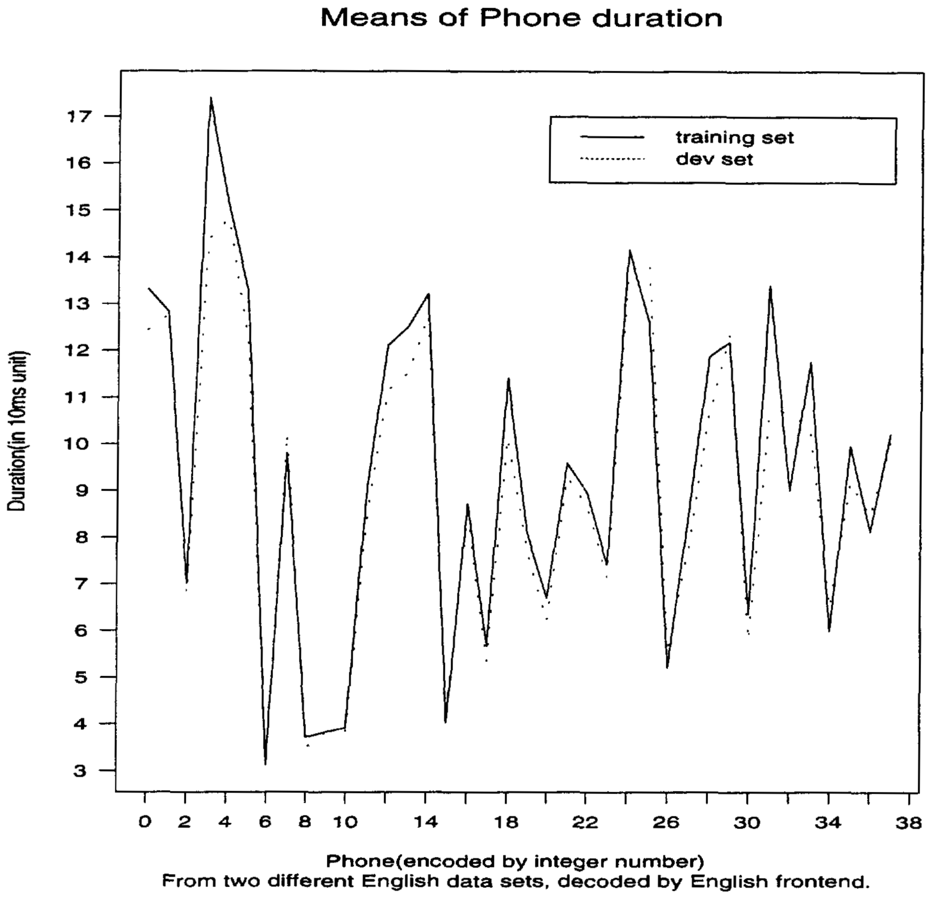


Figure 5.1: Analysis of Duration on the Same Language: English

preceding phone is a *vowel, fricative, stop, nasal, affricate or glide.* The duration models we used are given as:

$$P_D = \prod_{i=1}^{T}((1 - \alpha)P(O_i|O_i, O_{i-1} \in S) + \alpha P(O_i|O_i)) \tag{5.4}$$

where $P(O_i|O_i, O_{i-1} \in S)$ is the context-dependent model, and S is one of the six broad categories. $P(O_i|O_i)$ 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.

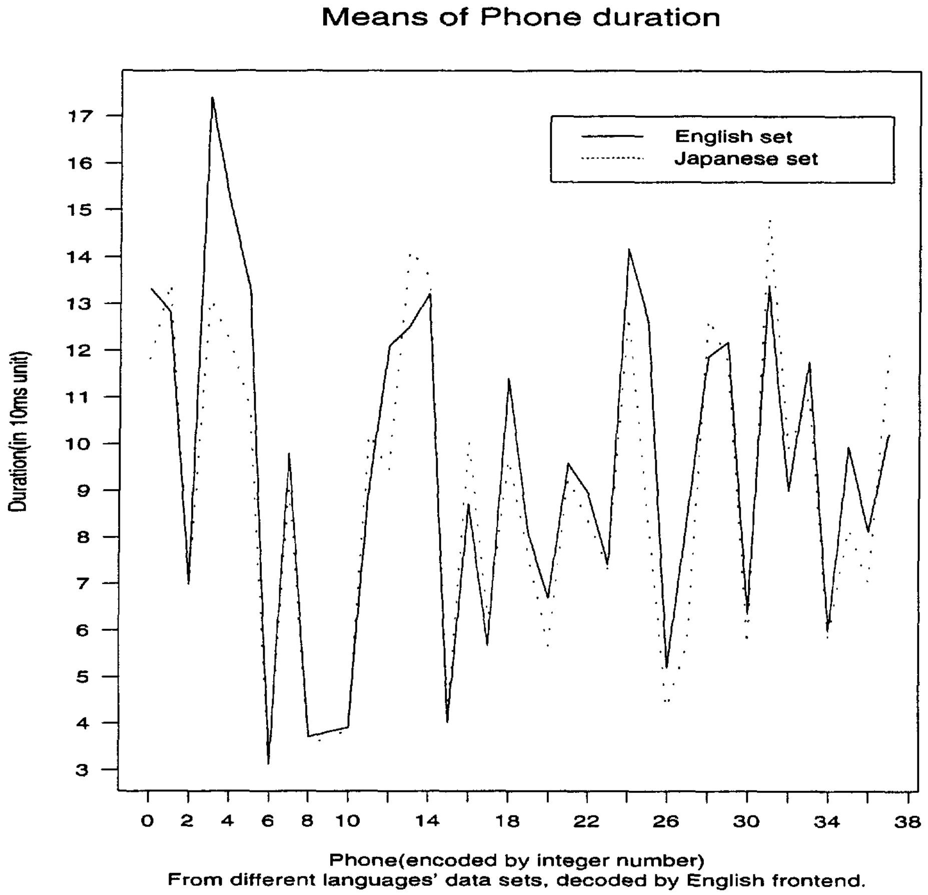


Figure 5.2: Analysis of Duration on Different Languages: English/Japanese

5.3 Neural-net Classifier

Recent research ([HZ94, YB95b]) has shown that simply combining all the information sources with different (linear) 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 propose to view the combination of scores as a classification problem rather than a combination of independent probabilities. The standard techniques thus amount to linear classification. 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 neural-network classification in many circumstances. Previously, the typical ways to combine multiple scores were either by some prior

Table 5.2: Phone Index in Table 5.1 and 5.2

Index	0	1	2	3	4	5	6	7	8	9
Phone	cl	iy	ih	ey	ae	eh	ah	uw	uh	ow
Index	10	11	12	13	14	15	16	17	18	19
Phone	aw	aa	ay	oy	er	aor	r	l	y	w
Index	20	21	22	23	24	25	26	27	28	29
Phone	hh	m	n	s	z	sh	th	dh	dx	f
Index	30	31	32	33	34	35	36	37	38	
Phone	v	ch	jh	p	t	k	b	d	g	

knowledge about the relative merit of different scores or by hill-climbing optimization. These techniques result is specific linear classifiers, and thus also suffer from limited modeling power.

In our experiment, neural-net based pattern recognizers are 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 ([BC89]). Figure 5.3 illustrates how the neural network is configured in the system.

5.4 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. We proposed a new way to dynamically optimize the LID models to increase the distinction between different

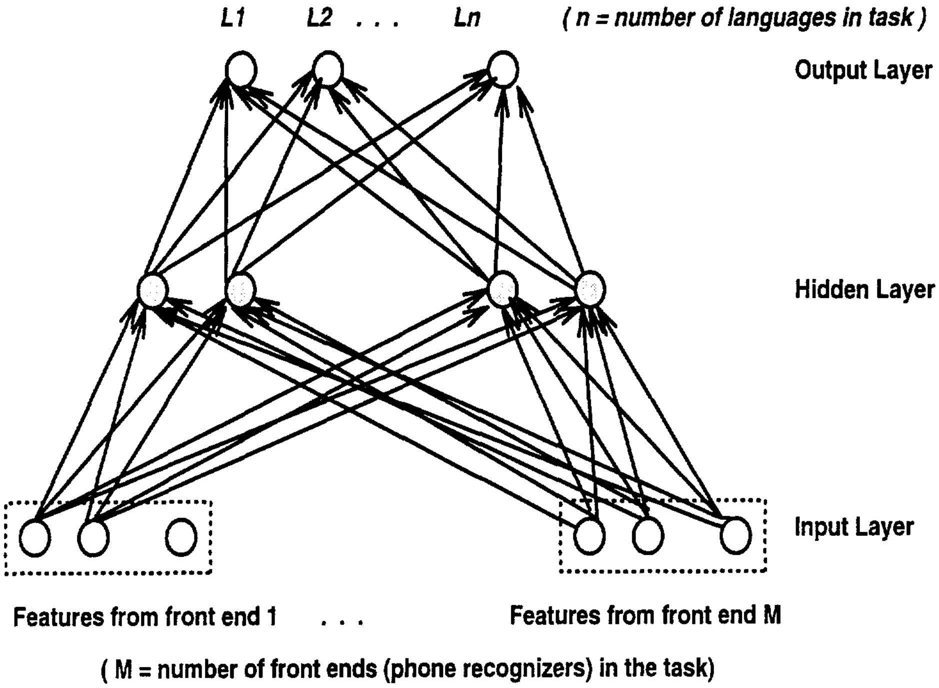

L1
L2 . . .
Ln
(n = number of languages in task)
Output Layer
Hidden Layer
Input Layer
Features from front end 1
. . .
Features from front end M
(M = number of front ends (phone recognizers) in the task)

Figure 5.3: Neural network as the final classifier

languages.

5.4.1 Cost function

In this section, we present our attempt to optimize the language model, so as to improve robustness with respect to the front end errors.

Our approach uses a cost function based on the maximum-likelihood LID scores calculated by the score generator. For any score generator in the system, the cost function for a particular input utterance is defined as:

$$E = (S_T - S_M)^2 \tag{5.5}$$

where

$$S_M = \max_i S_i, \qquad i = 1, ..., N \tag{5.6}$$

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:

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

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

Here we give an example to illustrate how the optimization is realized; this example is for the optimization of the forward language model; namely, optimization of the bigram and the unigram terms. Other LID models used in the system can be optimized similarly.

By using (4.1) and (5.5), for the forward language model,

$$\begin{aligned} E &= (S_T - S_M)^2 && (5.7) \\ &= (\sum_{i=1}^{T} log(\alpha P(O_i|O_{i-1}) + \beta P(O_i)) - S_M)^2 && (5.8) \end{aligned}$$

$$\begin{aligned} \frac{\partial E}{\partial P(O_i|O_{i-1})} &= 2(S_T - S_M)\frac{\partial S_T}{\partial P(O_i|O_{i-1})} && (5.9) \\ &= 2(S_T - S_M)\frac{\alpha}{\alpha P(O_i|O_{i-1}) + \beta P(O_i)} && (5.10) \end{aligned}$$

$$\begin{aligned} \frac{\partial E}{\partial P(O_i)} &= 2(S_T - S_M)\frac{\partial S_T}{\partial P(O_i)} && (5.11) \\ &= 2(S_T - S_M)\frac{\beta}{\alpha P(O_i|O_{i-1}) + \beta P(O_i)} && (5.12) \end{aligned}$$

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

$$\Delta P(O_i|O_{i-1}) = -\eta \frac{\partial E}{\partial P(O_i|O_{i-1})} \qquad (5.13)$$

update $P(O_i)$ by

$$\Delta P(O_i) = -\eta \frac{\partial E}{\partial P(O_i)} \qquad (5.14)$$

where η is the learning rate. *i.e.,*

$$P_{new}(O_i|O_{i-1}) = P(O_i|O_{i-1}) + \Delta P(O_i|O_{i-1}) \tag{5.15}$$

$$P_{new}(O_i) = P(O_i) + \Delta P(O_i) \tag{5.16}$$

After one iteration, the re-normalization of the bigram probabilities results in new bigram and unigram terms for the language models. 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 stopping criterion.

The advantage of this back-propagation based optimization is that it dynamically enhances the discrimination among the languages. The resulting bigrams (or unigrams) are 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.

5.4.2 Optimization Procedure

The actual optimization steps are illustrated in Figure 5.4.

An important issue in using this optimization method is how to prevent the models 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 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

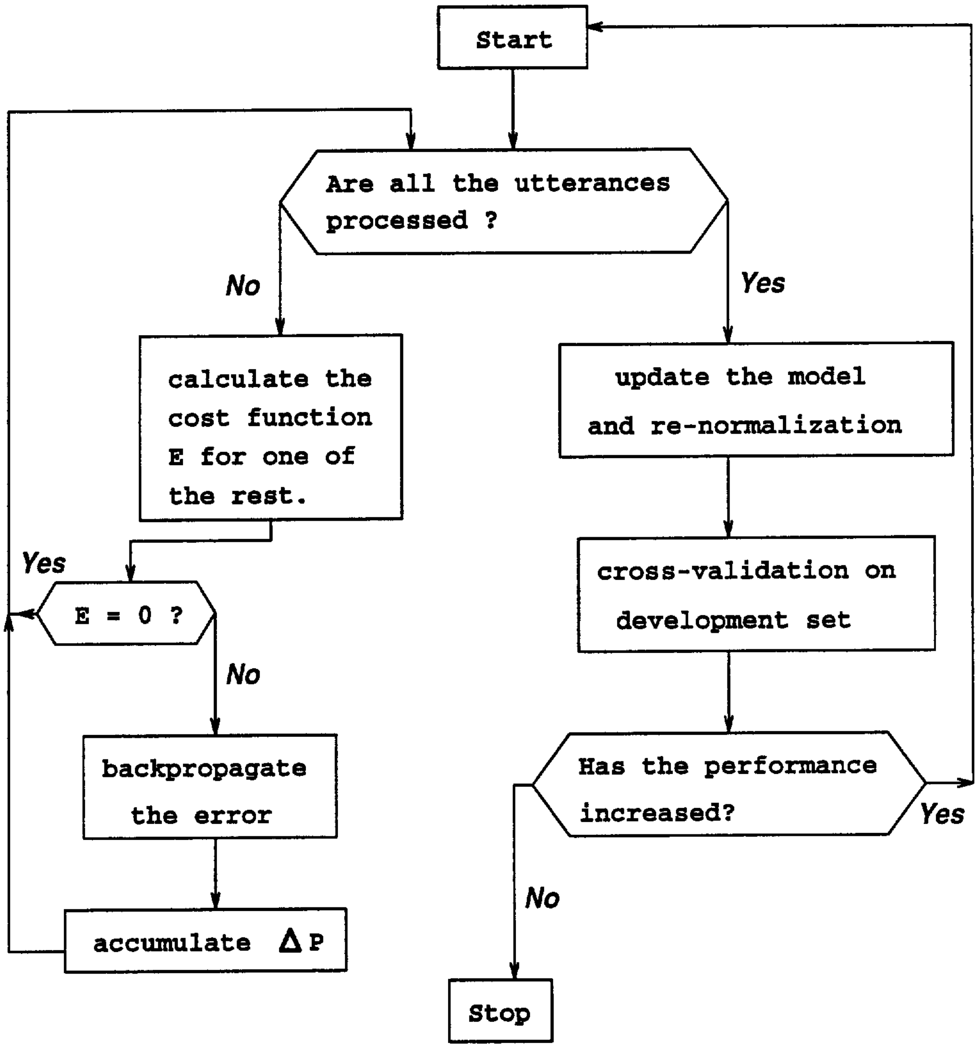

Start
Are all the utterances processed ?
No
Yes
calculate the cost function E for one of the rest.
update the model and re-normalization
Yes
E = 0 ?
No
cross-validation on development set
backpropagate the error
Has the performance increased?
Yes
No
accumulate ΔP
Stop

Figure 5.4: Optimization of the LID model based on back-propagation

needed were accumulated. The actual updating took place after all the utterances were processed once). (3) A small learning rate was used.

5.5 Robust Speech Signal Processing

For approaches to language identification based on phone recognition, how to increase the phone recognition robustness in different environments is critical to system robustness. For the approach presented in this dissertation, the LID models are trained exclusively through the output of the phone recognizers, handling the communication channel effect is especially important.

Since robust signal processing is not a major concern (or task) of this dissertation, we only implemented existing algorithms. In particular, two techniques were evaluated: RASTA[HMH93] and cepstral mean subtraction[Ata74, DM80].

5.5.1 RASTA Processing

RASTA (RelAtive SpecTrA) is a filtering technique based on the assumption that the temporal properties of the communication environment are quite different from the temporal properties of speech. The changing rate of nonlinguistic components in speech therefore often lies outside the typical dynamic rate of the vocal tract. The RASTA technique takes advantage of this fact and by band-pass filtering the signal, the effects caused by linear microphone characteristics (convolutional component in the signal in time domain) and additive noise can be removed partially.

In our experiment, *J-Rasta*[HMH93] was implemented.

5.5.2 Cepstral Mean Subtraction

This method is based on the assumption that the frequency characteristics of a communication channel are often fixed or slowly changing. By subtracting the long term average from the the logarithmic spectrum of the signal, the resulting signal representation is less sensitive to the environment. This is also called blind deconvolution of signals. One advantage of this technique is the simplicity of implementing it.

In our experiment, the cepstral mean of a whole test utterance is subtracted from each cepstral vector of the utterance.

Chapter 6

Experiments with the Proposed Improvements

In this chapter, we present our experiments with the improvements proposed in this dissertation. The front end (phone recognizers) developed for the baseline system was used throughout. First we present the evaluation of the two LID models within the structure of the baseline system. Then we present the comparison between the baseline system and the baseline system enhanced by each individual method of improvement. Finally we present the results when all these methods are used. Experimental results on channel normalization are also reported.

6.1 Experiments with the Proposed LID Models

In order to measure the gain achieved by using the proposed models, three experiments were carried out.

1. Experiment with the backward bigram language model (B)
2. Experiment with the enhanced duration model (ED)
3. Experiment with the backward language model and the enhanced duration model (B_ED)

These experiments are similar to those presented in the baseline evaluation (Chapter 4). They provide a quantitative measurement of the system performance when these models are used in isolation.

Table 6.1: New model evaluation

Task		B	ED	B_ED
Six Language	45-second	83.0%	57.1%	85.1%
	10-second	74.1%	47.0%	76.2%
Eleven Language	45-second	74.4%	44.1%	76.2%
	10-second	64.3%	35.0%	66.6%

The data sets used for training, developing and testing are the same as those used in the baseline system evaluation. All six phone recognizers are used in these experiments. Results are given in Table 6.1.

We see that the enhanced duration models alone are still inadequate. However, as an additional set of features, they improve the correct rates in both tasks.

6.2 LID Model Evaluation: Role of Phone Recognizer

Similar to the evaluation conducted in section 4.2.2, we also tested the backward language model and the combined LID model in systems with each phone recognizer individually. These experiments detail why several information sources and several phone recognizers are important to the performance of a LID system.

All the experiments were conducted on the eleven-language task.

6.2.1 Evaluation of the Backward Language Model

In this set of experiments, the system configuration is the same as the baseline system, except that only the backward bigram language models are used in the score generators. Each phone recognizer was used in turn in the system. The results of each score generator using the backward language models for each phone recognizer are given in Figure 6.1.

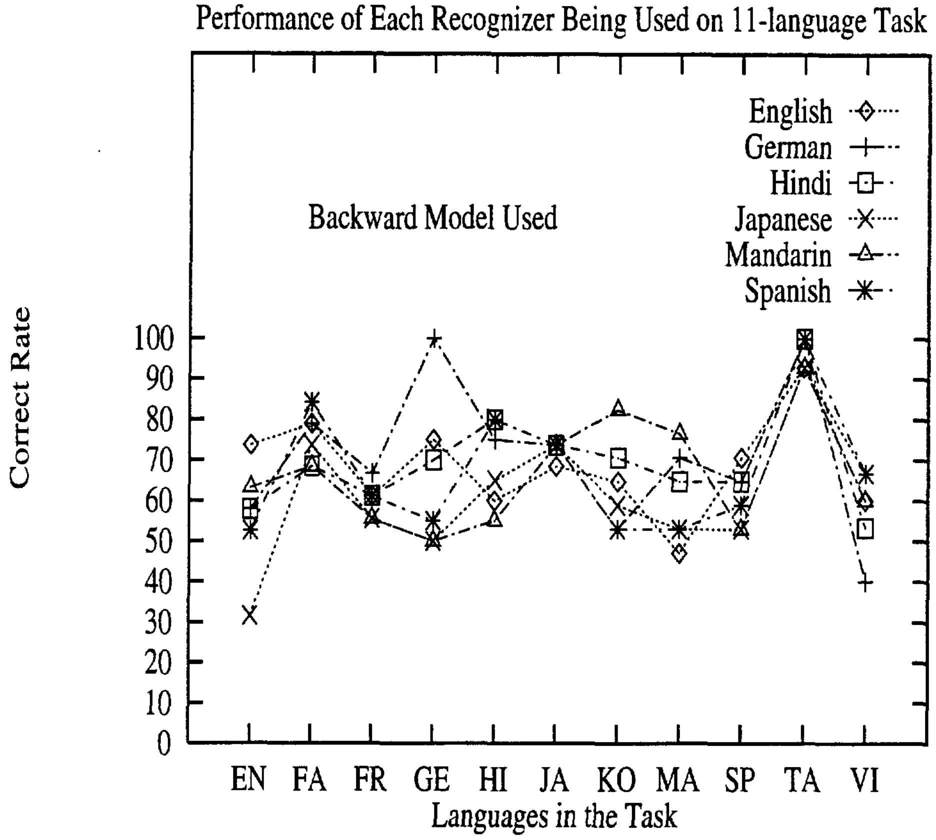


Figure 6.1: LID performance by each recognizer based on backward language model

Table 6.2: Overall performance of system: using the Backward language model

Recognizer	English	German	Hindi	Japanese	Mandarin	Spanish	ALL_SIX
Mean	68.2%	71.3%	69.2%	60.5%	65.6%	66.7%	78.97%
S.D.	11.5%	17.3%	12.0%	15.2%	13.0%	14.9%	9.4%

The LID results for each language with different phone recognizers are different; no single system based on one specific phone recognizer could consistently outperform systems based on the other phone recognizers on all the languages in the task. Figure 4.2 and Figure 6.1 confirm the necessity of using multiple phone recognizers in one LID system. More details can also be found in Table 6.2.

6.2.2 Evaluation of the Combined Language Model

Although linear combination is not always a good way to integrate multiple scores from different LID models, for simplicity (to avoid the training of a classifier), we used this simple method to combine all the three models. The combined LID model is:

$$P_{LID} = \prod_{i=1}^{T}(\alpha P_l(O_i|O_{i-1}) + \beta P_l(O_i|O_{i+1}) + \gamma P_d(O_i)|O_i, O_{i-1} \in S) \qquad (6.1)$$

where, P_l is the language model term, and P_d is our duration model. In this experiment, α is set to 1, β is set to 0.5 and γ is set to 0.1.

The combined models were evaluated within the same system architecture as used for evaluation of the forward and backward language models. The results for the combined LID models are given in Figure 6.2. More details can be found in Table 6.3.

Table 6.3: Overall performance of system: using the Combined language model

Recognizer	English	German	Hindi	Japanese	Mandarin	Spanish	ALL_SIX
Mean	76.4%	75.4%	73.9%	68.7%	65.1%	70.3%	83.1%
S.D.	10.6%	14.7%	11.1%	13.1%	13.2%	12.3%	9.4%

Compared with the results given in Table 4.3 and Table 6.2, we see that:

- The forward language model is slightly more important than the backward language model. This conclusion is consistent with our results from model evaluation with the complete baseline system (in which all the phone recognizers are used).

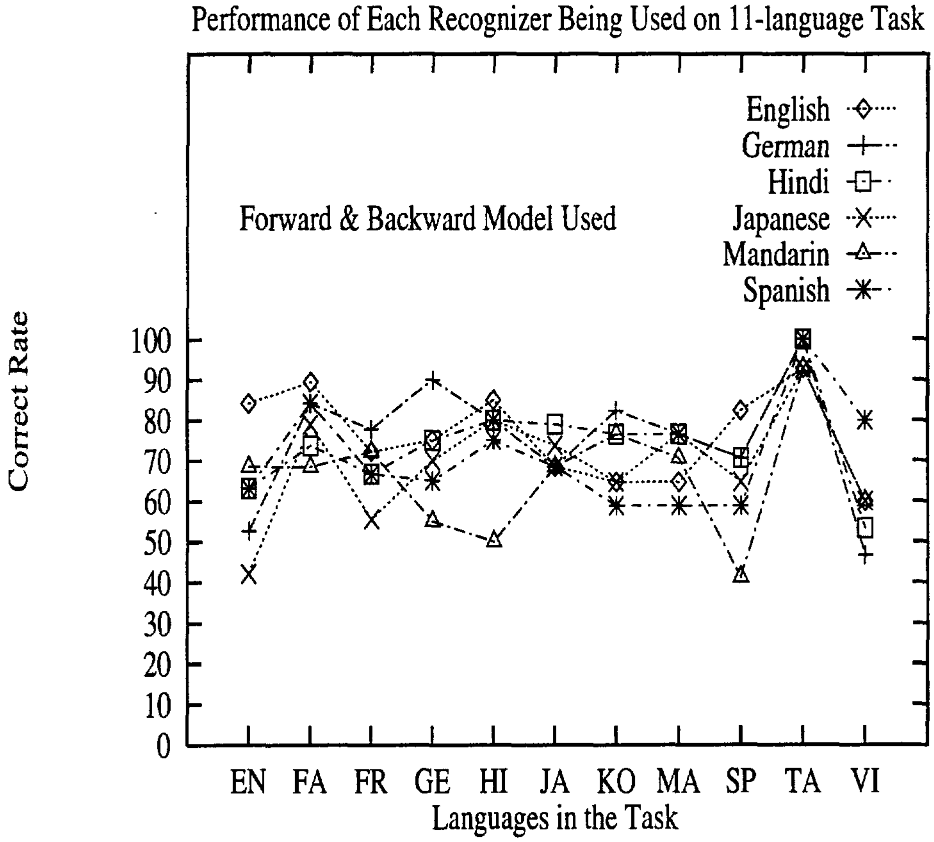


Figure 6.2: LID performance by each recognizer based on the combined language model

- The combined model can outperform any single model. Again, this is consistent with the conclusion of the whole system evaluation.

6.2.3 Experiment with the Complete Front End

Comparative experiments based on the complete front end (all six phone recognizers are used) were also conducted. The results are shown in Figure 6.3.

The overall performance of the systems using different models in the score generator is summarized in Table 6.4. Note: these results are obtained by using (4.3). Even using a simple final classifier, the system with several information sources outperforms the system with only one information source.

For comparison, the performance of systems with each of the six phone recognizers as

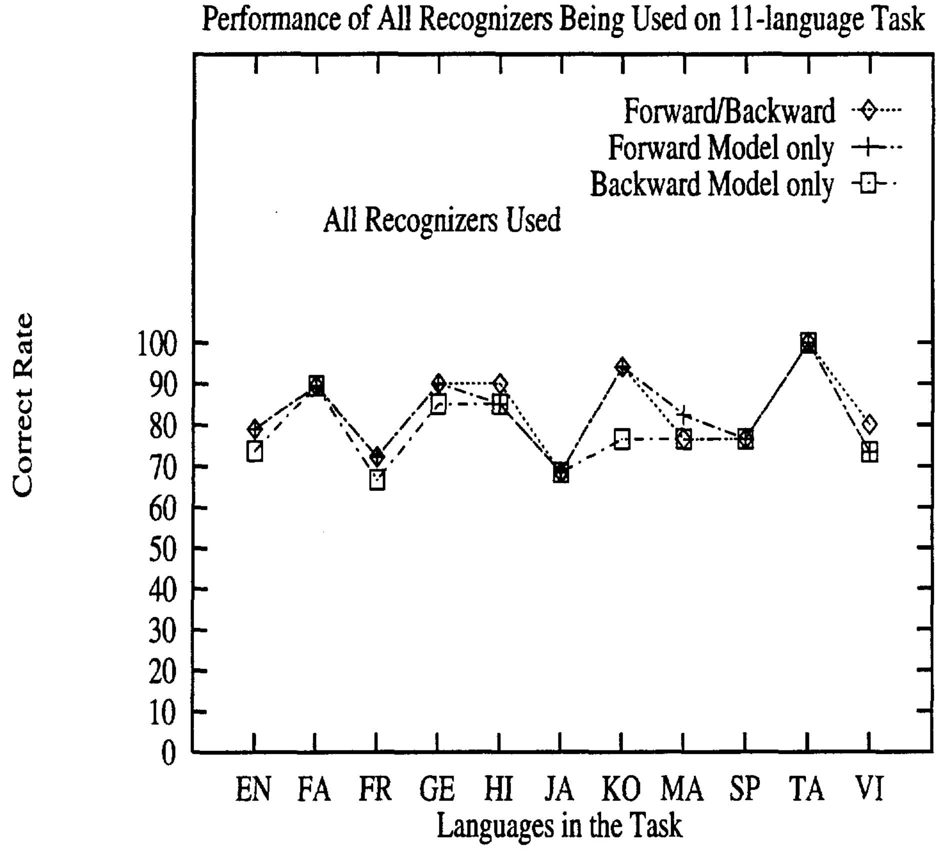


Figure 6.3: LID performance of different models using the complete front end

the front end using different language models is summarized in Figure 6.4.

6.3 Experiments with the Baseline System Enhanced by Each Method

In order to compare the impact of each proposed method on the baseline system, we conducted experiments with the baseline system enhanced by each method individually.

6.3.1 Baseline System Plus Backward Bigram Language Model

In these experiments, in addition to the two models used in the baseline system, the backward bigram LID models are also used as the third set of models to enhance the

Table 6.4: Overall performance of systems based on different models: Using all six recognizers and (4.3)

Language Model	Forward	Backward	Combined
Mean	81.6%	79.0%	83.1%
S.D.	9.5%	9.4%	9.5%

modeling accuracy. The data used to train the backward bigram language models were the same as those used to train the language models in the baseline system.

Compared with the baseline system, adding the backward language models doubles the memory requirement for the language models, while the increase in computational time (CPU cycles) during testing is almost negligible.

The resulting enhanced system was evaluated on the six- and eleven-language tasks. Results are summarized in Table 6.7 (labeled as BS_LAN).

6.3.2 Baseline System with the Enhanced Duration Model

In these experiments, the duration models in the baseline system were replaced by the enhanced duration models. The new duration models were trained with the data used to train the baseline duration model.

The enhanced version of the baseline system was evaluated on the six- and eleven-language tasks; results are summarized in Table 6.7 (labeled as BS_DUR). The improvement achieved by the new duration model is minor. The duration information clearly depends not only on a small context (such as contiguous phones); how to robustly model duration information is still an interesting issue in language identification research.

6.3.3 Baseline System with a neural network as the Final Classifier

In these experiments, the linear classifiers are replaced by feed-forward networks[YB95d]. The data used to train the linear classifiers are used to train the neural networks.

Various neural network architectures (different numbers of hidden nodes) were tested on

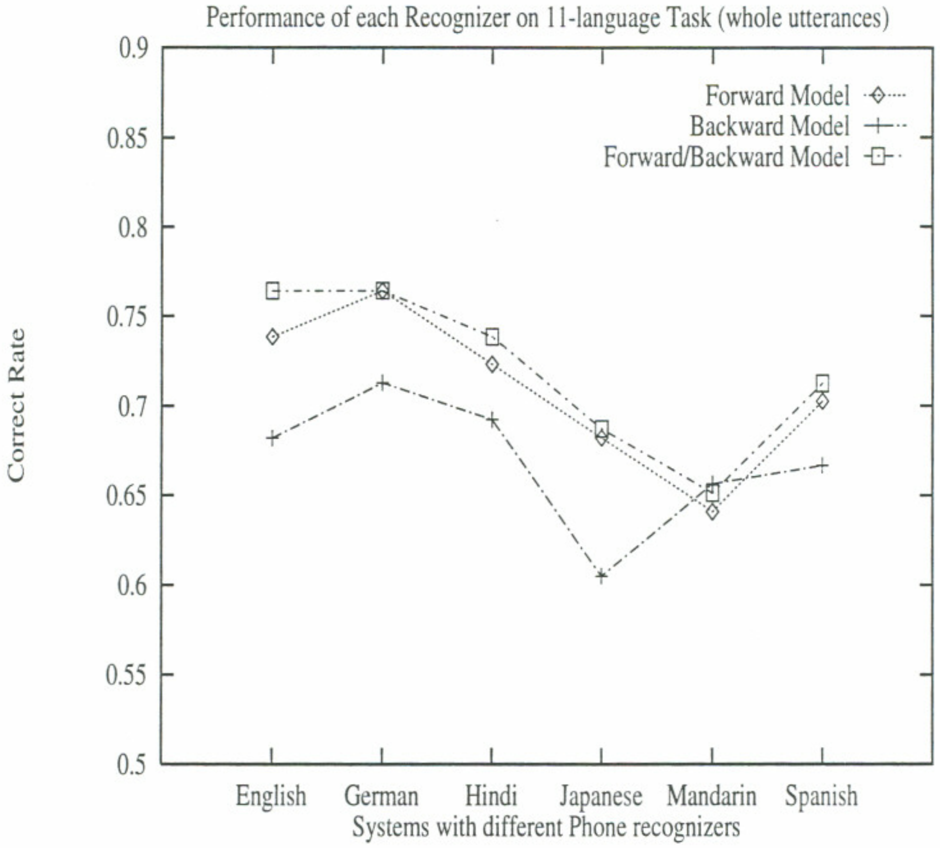


Figure 6.4: Overall LID performance by systems with different front ends

the development set. The best architecture (with 20 to 30 hidden nodes for different tasks) was used to do the final training and testing.

In order to compare the performance of the neural network and linear classifier as the final classifier in the system, the following two sets of experiments were conducted:

- Only one set of features (forward language model) was used in the score generator.
- All three sets (forward and backward language models, duration models) of features were used in score generator.

All the experiments were conducted on both the six-language and eleven-language tasks. The dimensions of the LID score vector being sent to the final classifier in these experiments are given in Table 6.5.

Table 6.5: Size of the score vector: Input to the final classifier

Experiment	Dimension($M \times N \times L$)
SIX_S	$6 \times 6 \times 1 = 36$
SIX_M	$6 \times 6 \times 3 = 108$
11_S	$6 \times 11 \times 1 = 66$
11_M	$6 \times 11 \times 3 = 198$

Results for these experiments are given in Table 6.6, where **SIX_** and **11_** denote six- and eleven-language tasks respectively, **_L** and **_N** denote linear and neural-net classifiers respectively, and **_S** and **_M** distinguish between the single and multiple sets of features. For comparison, the results of **_L_M** are also listed in Table 6.7 (labeled as BS_NN).

6.3.4 Baseline System with Optimization

In these experiments, the language models are optimized by the formula given in Chapter 5. 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 effect of optimization on the bigram terms in the forward language model is shown in Figure 6.5; the effect on the final results is shown in Figure 6.6. In Figure 6.5, 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 6.6, for the English Score generator, the performance on data from English, Hindi, Japanese and Spanish are improved by the optimization.

Table 6.6: Results (correct rate) for all the experiments

Approach	45-sec. utterance	10-sec. utterance
SIX_L_S	87.5%	76.2%
SIX_N_S	88.4%	77.3%
SIX_L_M	90.2%	79.2%
SIX_N_M	92.0%	81.6%
11_L_S	80.5%	69.1%
11_N_S	82.6%	70.0%
11_L_M	83.6%	70.1%
11_N_M	86.7%	73.8%

Table 6.7: LID Results: Baseline System (BS) Enhanced by Each Method

	Six-language Task		Eleven-language Task	
Enhancement	45-second	10-second	45-second	10-second
BS	86.1%	76.2%	78.5%	67.0%
BS_LAN	88.4%	78.9%	82.6%	70.1%
BS_DUR	87.5%	76.2%	79.0%	68.0%
BS_NN	90.2%	79.2%	83.6%	70.1%
BS_OPT	89.3%	79.7%	83.1%	71.4%

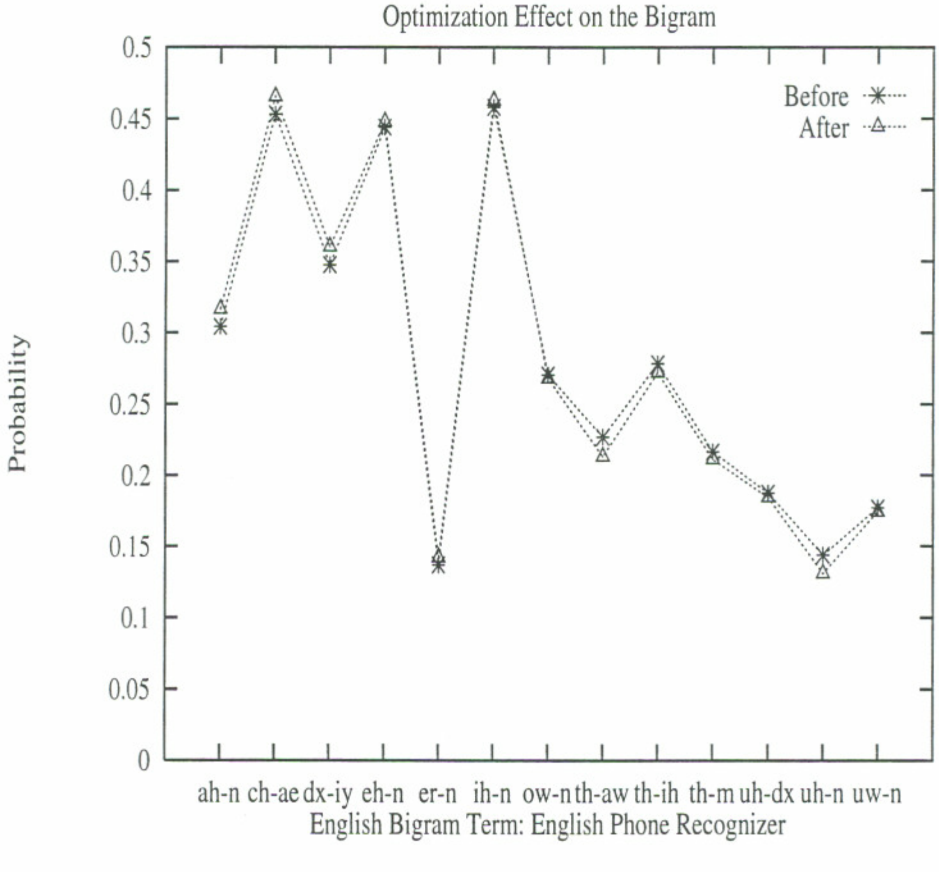


Figure 6.5: Analysis of Language Model (English Front End)

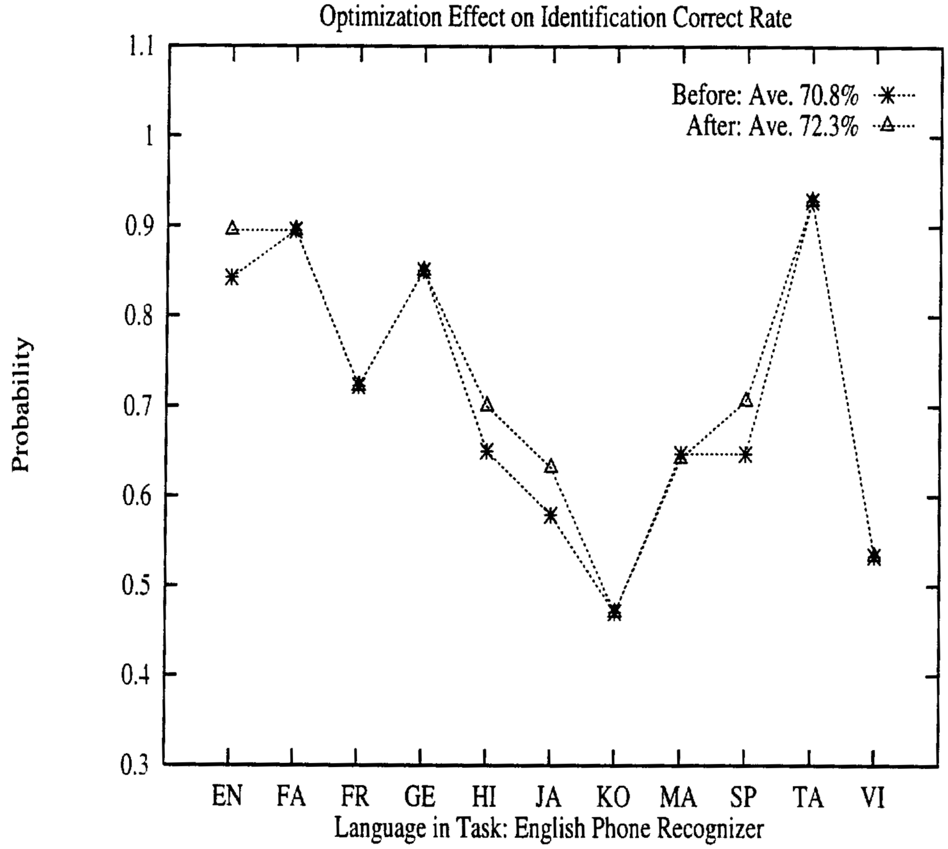


Figure 6.6: Optimization Results (English Front End)

The results of the system after optimization are summarized in Table 6.7 (labeled as *BS_OPT*).

6.4 Combining the Enhancements

In this section, we present the experiments with the improved system, with all enhancements applied to the baseline system. The front end (the six language-dependent recognizers) are the same as the front end in our baseline system. The score generator has three sets of LID models: the forward language model, the backward language model and the enhanced duration model. Both language models are optimized iteratively, and the final classifier is the neural network studied. The final system was evaluated on the six-, eleven- and nine-language tasks.

6.4.1 Evaluation on the Six- and Eleven-language Tasks

The final enhanced baseline system is evaluated on the six- and eleven-language tasks for comparison purpose. The results are summarized in Table 6.8. For comparison, the results of baseline system are also listed.

Table 6.8: Language Identification Correct Rate: the Enhanced system

	Task	Six-language Task	Eleven-language Task
Baseline System	45-second	86.6%	78.5%
	10-second	76.2%	67.0%
Final System	45-second	92.0%	86.7%
	10-second	81.6%	73.8%

The confusion matrices for the best system on the six-language task are given in Table 6.9 and Table 6.10. The confusion matrices for the best system on the eleven-language task are given in Table 6.11 and Table 6.12. The rows of the matrices correspond to the languages actually being spoken and the columns indicate the languages identified.

Table 6.9: Confusion matrix for the 45-second long utterances: Six-language Task

Language	English	German	Hindi	Japanese	Mandarin	Spanish
English	17	0	1	0	0	1
German	0	18	0	0	0	2
Hindi	1	1	18	0	0	0
Japanese	0	0	0	18	1	0
Mandarin	0	0	0	1	16	0
Spanish	0	0	1	1	0	16

Table 6.10: Confusion matrix for the 10-second long utterances: Six-language Task

Language	English	German	Hindi	Japanese	Mandarin	Spanish
English	60	0	5	1	0	3
German	3	51	3	1	2	5
Hindi	2	1	53	5	1	3
Japanese	1	0	6	48	1	5
Mandarin	2	0	1	3	44	2
Spanish	0	0	8	2	2	46

Table 6.11: Confusion matrix for the 45-second long utterances: 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

We see 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 caused by the language-dependent phone recognizer, which uses forced-choice best-path searching. 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. Also, imperfect performance of the recognizer, no explicit model of acoustic information and not strictly balanced sizes of training data for languages could be other major reasons.

Table 6.12: Confusion matrix for the 10-second long utterances: 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

6.4.2 Comprehensive Evaluation on the Nine-language Task

As a final test of system performance, new data (which are not involved in the system development period) are needed. We reserved the nine-language task data for this purpose. All the results reported in this section are obtained by running the final system on the data for the first time.

As mentioned in Chapter 3, the data for the nine-language task were collected via digital telephone lines (T-1 lines), which are different from the training and testing data used before. A simple zero energy detection algorithm is implemented to delete contiguous zeroes in the speech wave files to avoid the mathematical exceptions during LPC analysis.

The system was evaluated on the tasks specified by NIST in their 1995 March evaluation, which includes:

- six-language closed-set test.
- six-language open-set test.
- nine-language closed-set test.
- nine-language open-set test.
- EN-L pair test.

Details of these tasks can be found in Chapter 3.2.3.

The closed-set test systems are identical to all the systems presented before; for open-set test systems, one more output node (background-language node), representing the languages which were not in the task (background language), is added to the final classifier (neural network) of the system. The corresponding LID models for the background languages were created by averaging all the models of the languages in the task.

In the open-set test, it is not assumed that the set of training data bears any specific relation to the background languages. In the NIST'95 data, the background languages for the six-language task are: Czech, French, Portuguese, Tamil and Vietnamese; two of them (Czech and Portuguese) were not used in training. The background languages for the 9-language task are Czech and Portuguese; the system was never trained on the data from these two languages.

All the training data used were the data used to train the system for the eleven-language task presented in previous section. For the closed-set system, only the data from the languages in the task are used. For the open-set test system, all the data are used to train the neural-net based final classifier, with all the data from the languages not in the task labeled as background for the training of the background-language node.

The results for the six- and nine-language tasks are summarized in Table 6.13 and 6.14.

Table 6.13: Language Identification Correct Rate: Six-language Task

Task	Closed Set	Open Set
Whole	93.3%	82.7%
10-Second	81.1%	65.0%

Table 6.14: Language Identification Correct Rate: Nine-language Task

Task	Closed Set	Open Set
Whole	87.8%	72.7%
10-Second	74.0%	62.1%

The confusion matrices for the open-set tasks are given in Table 6.15, 6.16, 6.17 and 6.18.

Table 6.15: Confusion matrix for the 45-second long utterances: Six-language Task

Language	English	German	Hindi	Japanese	Mandarin	Spanish	Other
English	15	2	0	0	0	0	3
German	1	16	0	0	0	0	3
Hindi	0	0	15	0	0	1	4
Japanese	0	0	0	18	0	0	2
Mandarin	0	0	0	0	16	0	3
Spanish	0	0	1	1	0	15	6
Other	1	2	5	0	0	3	89

Table 6.16: Confusion matrix for the 10-second long utterances: Six-language Task

Language	English	German	Hindi	Japanese	Mandarin	Spanish	Other
English	59	6	0	0	2	0	11
German	3	53	1	0	0	0	16
Hindi	1	3	29	0	0	5	33
Japanese	1	0	1	43	1	1	27
Mandarin	1	6	2	1	40	1	22
Spanish	2	0	2	1	1	26	38
Other	21	16	19	16	7	12	271

Table 6.17: Confusion matrix for the 45-second long utterances: Nine-language Task

Language	EN	FR	GE	HI	JA	MA	SP	TA	VI	OT
EN	18	0	1	0	0	0	0	0	1	0
FR	0	15	0	0	1	0	0	0	0	4
GE	0	0	16	1	0	0	0	0	0	3
HI	0	1	0	16	1	0	0	0	1	1
JA	0	0	0	0	18	0	0	0	2	0
MA	0	1	1	0	0	15	0	0	2	1
SP	0	3	0	0	0	0	17	0	0	0
TA	0	0	0	0	1	0	0	19	0	0
VI	0	0	0	1	0	0	0	0	19	0
OT	2	2	2	10	4	1	11	0	0	7

Table 6.18: Confusion matrix for the 10-second long utterances: Nine-language Task

Language	EN	FR	GE	HI	JA	MA	SP	TA	VI	OT
EN	62	0	5	2	0	3	0	0	6	0
FR	0	52	2	4	2	2	3	1	3	8
GE	4	6	51	3	0	2	0	0	1	6
HI	2	3	3	49	0	0	3	2	2	7
JA	0	2	1	3	53	1	2	2	8	2
MA	1	5	5	4	0	47	2	0	2	7
SP	1	9	1	9	1	2	41	3	1	2
TA	0	0	0	3	3	0	0	63	0	2
VI	0	4	0	6	1	0	0	1	44	10
OT	5	12	17	28	21	7	13	3	7	35

The performance decrease on the open sets (compared with the corresponding close set tests) is mainly caused by false detection of the background languages (in the matrices, they are labeled as *OT*). For the six-language open-set task, since the data of the background languages used to train the neural network include data from five languages, the performance is relatively better than that of the nine-language open-set test. Within our current statistical modeling paradigm, rejection of a background language with no training data at all is an extremely challenging task. The other important reason for better performance on the six-language task is that data from three of the five background languages (French, Tamil and Vietnamese) were available. We thus conclude that this task is manageable if training data are available.

As is traditional in language-identification system evaluation, we also evaluated our system on English-other language-pair identification. The results are given in Table 6.19.

Table 6.19: Results on the English-other language-pair task

Language	FR	GE	HI	JA	MA	SP	TA	VI	Average
45-second	100%	93%	97%	100%	100%	97%	100%	95%	98%
10-second	99%	91%	97%	99%	93%	96%	98%	92%	96%

6.5 Channel Normalization

As mentioned in Chapter 5, although robust signal processing is not a major concern of this dissertation, it is very important for system performance. We experimentally implemented the J-RASTA and the cepstral mean subtraction techniques. Initial results showed that the improvement on phone recognition accuracy achieved by cepstral mean subtraction was larger than that achieved by RASTA, so cepstral mean subtraction is used in our system[YB95a].

6.5.1 Impact on the Phone Accuracy

After the LPC_Cepstral coefficients are calculated, the mean of the cepstral coefficient vectors of the entire testing utterance is subtracted from each cepstral vector. We retrained our phone recognizers using the same HMM topology, rebuilt our LID system with the same sets of models, and estimated model parameters with the same training routines using the same training set and test set used before. The phone recognition results are given in Table 6.20. For comparison, previous phone accuracies are also listed.

Table 6.20: Impact of Channel Normalization on Phone Accuracy

Accuracy	Eng	Gem	Hin	Jap	Man	Spa
Before	46.8%	46.6%	48.1%	56.3%	36.3%	54.6%
After	50.4%	47.2%	51.1%	57.4%	43.4%	56.7%

After applying channel normalization, the performance of all six phone recognizers improves, though to varying degrees. For telephone speech, collected via different handsets and phone lines, this simple technique effectively increases the system robustness.

6.5.2 Impact on the LID Results

After retraining the phone recognizers with cepstral mean subtraction, we retrained our best system. The new system was evaluated on the nine-language and the eleven-language closed-set tests. Results are summarized in Table 6.21 and 6.22. Adding channel normalization effectively increases our system performance. The major drawback of our implementation is the long time delay: in order to calculate the mean vector, one needs to wait until the whole utterance has been processed. Although obvious approximations for real-time implementation exist, evaluating their effect on LID is an important task for future consideration.

Table 6.21: Impact of Channel Normalization on LID: Nine-Language Task

Error Rate	WHOLE	10-SECOND
Before	12.2%	26.0%
After	9.9%	22.5%

Table 6.22: Impact of Channel Normalization on LID: Eleven-Language Task

Error Rate	WHOLE	10-SECOND
Before	12.3%	26.3%
After	9.2%	22.9%

6.6 Summary of Results

All the relative improvements by each proposed methods are summarized in Table 6.23; the results are achieved on the eleven-language task. In the table, the error reduction is the average percentage of the error reductions achieved on the 45-second and 10-second long utterances. The statistical significance test used is the multinomial Chi-Square Test ([DW83]) with a 5% significance level.

Although the improvement achieved by each individual methods is not significant (given the limited amount test data), the final system is substantially better than the baseline approach.

Table 6.23: Summary of Results (Error rate): 11-language task. The final column lists whether the observed differences were statistically significant at the 5% level, assuming a multinomial distribution

	45-second	10-second	Error Reduction	Significance
Baseline	11.5%	33.0%	N/A	N/A
+ Backward (B)	17.4%	29.9%	15%	No
+ Duration (D)	21.0%	32.0%	3%	No
+ Neuralnet (N)	16.4%	28.3%	18%	No
+ Optimization (O)	16.9%	28.6%	18%	No
+ BDNO	12.3%	26.4%	35%	Yes
+ BDNO + Channel	9.9%	22.5%	43%	Yes

Chapter 7

Experiments with Conversational Speech and New Language Adaptation

For many potential applications of language identification, the system should be able to handle interactive speech and should have the ability to easily be familiarized with the characteristics of new languages which are not in the task (new language adaptation). Here we present our efforts in dealing with these new pursuits: processing conversational speech and adaptation to new languages. Due to the limited data which are publically available, only preliminary results are reported.

As observed in the previous experiments, in general the relative system performance is consistent on the long segment (whole utterance) and short segment (10-second utterance) tasks: The system which is better on the long-segment task also tends to be better on the short-segment task. In order to minimize the cost of the comparative experiments, only results on the long-utterance tasks are evaluated and reported in this chapter.

7.1 Language Identification using Conversational Speech

Identification of conversational speech is one of the latest interests for LID research. Compared with the processing of monologue data, conversational speech data present new challenges. For speech recognition, researchers have reported dramatic performance decreases when the systems are switched to process conversational speech. System performance on the ARPA Wall Journal corpus[PB92] and the Switchboard corpus[GHM92] tasks is a typical example. Impressive results have been achieved on monologue speech data on a fairly large task (for the ARPA Wall Street Journal task, the state-of-the-art system has an 8% word error rate on a task with 65k words[WLO+95]). The system developed for the Switchboard task with the same algorithms by the same research group still has a word error rate of around 50%[YWB94]. This reflects the challenge of conversational speech.

7.1.1 Analysis of the Conversational Speech

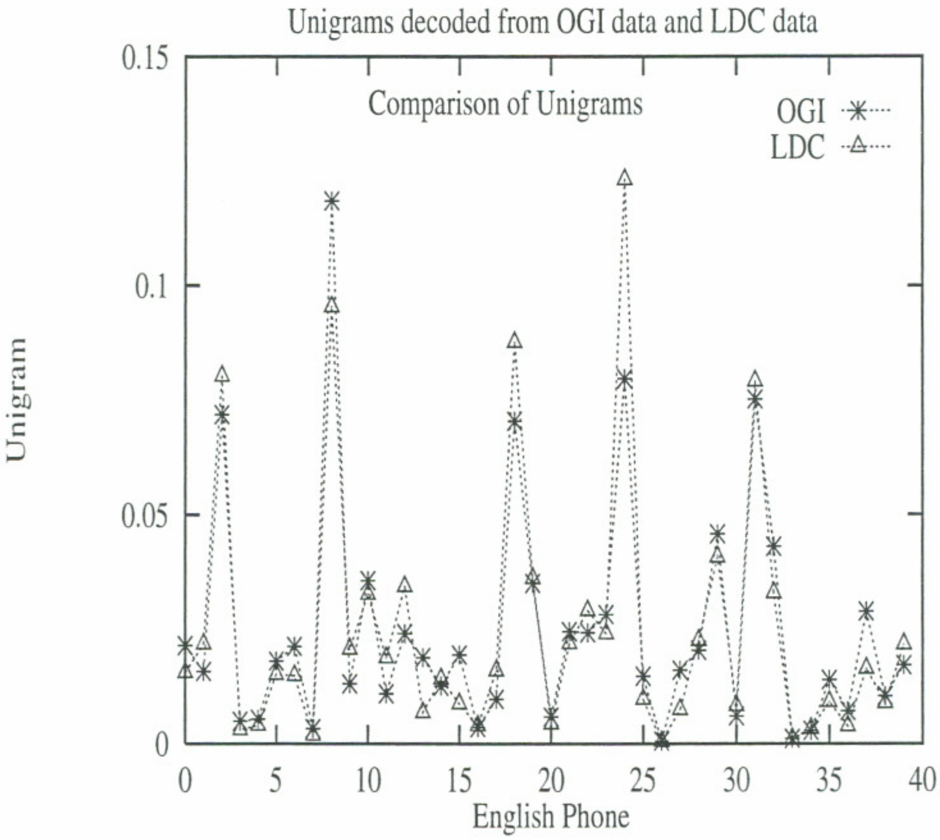


Figure 7.1: Phone distributions of monologue speech and conversational speech (by English Phone Recognizer)

Conversational speech is quite different from elicited monologues (whether read or spontaneous). Conversations contain frequent filled pauses (such as "uhuh"), repetitions, hesitations, excitations, false starts, and particularly poor articulation. As a result, conversational speech has a much higher variability in phone quality and duration. This causes problems when the recognizers trained on monologue speech data are used to recognize the underlying phones of a conversational utterance. By listening to some of the data from the LDC Call Friend sample release, we found that there are:

- more filled pause segments,
- more emotional and bigger prosodic variations,
- long silence segments.

Using the English recognizer developed for the system presented in previous chapters as an automatic labeler, the differences between the monologue and conversational speech data used in this study are illustrated in Figure 7.1 and 7.2. The phone index in these tables is given in Table 5.2.

Note the substantial differences that are observed. Using the same training data for the eleven-language task and the algorithm used to develop the system for monologue speech, we implemented a nine-language task system and evaluated it on the LDC Call Friend data. Compared with the results we achieved on the monologue data, the error rate is tripled (see Table 7.1).

7.1.2 Improving System Performance

As mentioned above, one major difference between the training data (monologue) and the testing data (conversational speech) is the presence of long silence segments in the latter. This presents a problem to the front end. In our implementation, channel normalization is implemented by cepstral mean subtraction, which does not consider if the frame being processed is speech or not. In our training data, the silence segments are generally only a small part of the data files. After subtraction, not only the channels, but also the speakers in the training data are thus normalized. With the long silence segments in conversational speech, the mean of a conversational utterance contains more information about the channel, so the speakers are not well normalized. This causes a mismatch between the training and testing environments.

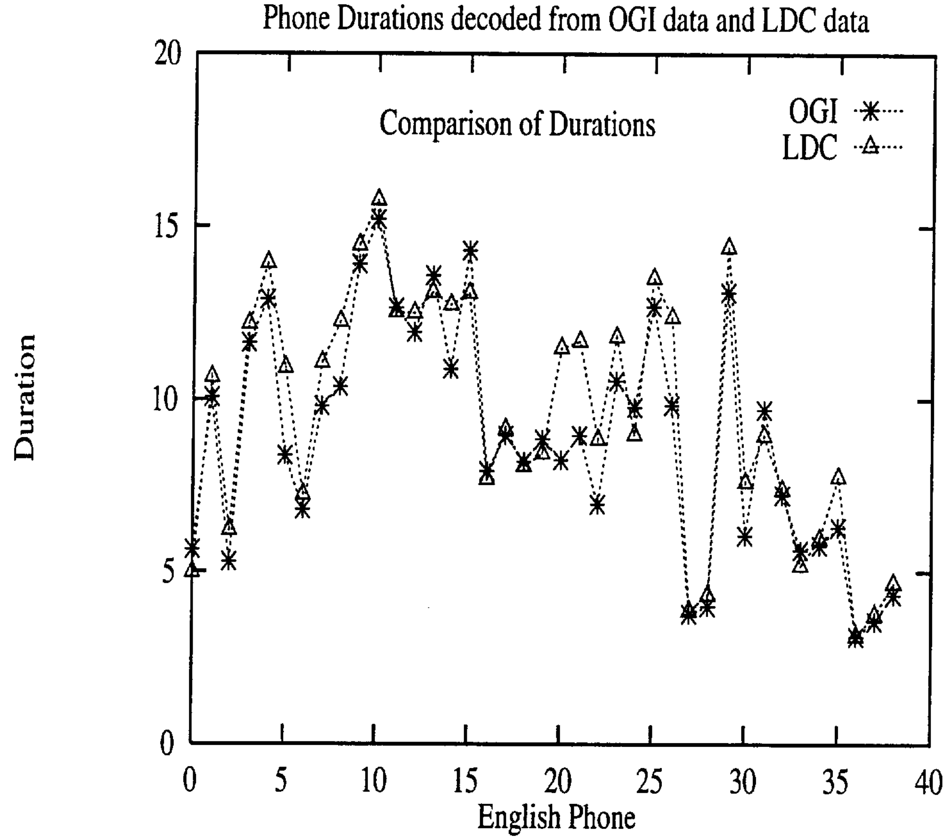


Figure 7.2: Distributions of phone durations for monologue speech and conversational speech (English Phone Recognizer)

In order to overcome this, we experimentally proposed a simple energy-based algorithm to delete the long silence segments. Within a segment of T frames, for frame i, the energy E_i is calculated, and compared with the threshold (TE) given in (7.1).

$$TE = \frac{1}{\alpha \times T} \times \sum_{i=1}^{T} E_i + \beta \times E_{min} \tag{7.1}$$

where,

$$E_{min} = \min_{i=1}^{T} E_i \tag{7.2}$$

The energy for each frame (with N samples) is calculated as:

$$E_i = \sum_{j=1}^{N} x_j^2 \tag{7.3}$$

where x_j is a sample in the frame.

Frame i is classified as:

$$Frame\ i\ is: \begin{cases} Speech, & \text{if TE} \leq \text{E}_\text{i} \\ Silence, & \text{if TE} > \text{E}_\text{i} \end{cases}$$

In our implementation, α is set to 1000.0 and β is set to 6.0.

By examining the outputs (decoded phone strings), we find there are some obvious error patterns in recognition results produced by the phone recognizers. A post-processing algorithm is proposed to correct some of these obvious mistakes. It performs the following corrections on the outputs of the phone recognizers. The rules were generated based on the analysis of the differences in the unigrams and bigrams between the monologue and conversational speech.

- single consonants surrounded by silences are not allowed.
- segments containing more than 4 contiguous consonants are not allowed.
- single (or repeated) nasal/(z,ah,s) pairs are not allowed.
- contiguous silence segments are merged.

The processed phone strings are sent to the score generators.

7.1.3 Experiments and Results

Three experiments are carried out in order to measure the improvement achieved by the pre-process (silence detection) and post-process (forced correction).

1. system implemented with only pre-processing.
2. system implemented with only post-processing.
3. system implemented with both pre-processing and post-processing.

All these systems were trained with the same data sets used before (from the OGI_TS database), and evaluated on the Call Friend data. Results are given in Table 7.1. For comparison, the result of the monologue system is also listed in the table. The error reductions achieved (compared with the monologue system) are also listed.

Table 7.1: LID Results on conversational speech

Approach	Result	Error Reduction
Monologue system	59.7%	N/A
Pre-process system	69.4%	24.1%
Post-process system	70.8%	27.5%
Combined system	76.4%	41.4%

Both pre- and post-processing cut the system error rate by approximately a quarter. Although the combined system achieves an error reduction of more than 40%, the performance of this system is still inferior to the performance we achieved on the same task with monologue speech as input. This may suggest that the task for conversational speech is inherently more difficult; however, training the system on conversational speech data may be sufficient to recover the results obtained with monologues. (This requires the availability of conversational speech databases.)

7.2 New Language Adaptation

The training of our current system needs large amounts of data from each language in the task, and it is possible to build a system only when sufficient data from all the languages are available. For many LID applications, an iterative method of training data collection and system development would be preferable: with a limited amount of data, building a system first and during the real world application, collecting more data to refine the system.

The fact that we used a fixed set of phonetic front ends is helpful in this regard, since this stage does not need to be retrained for new languages.

In order to achieve this, the requirement for training data by the LID algorithm should be relaxed while still maintaining decent system performance. In this section, we present our efforts in dealing with the problem of adaptation to new languages. The data used for this study are the monologue data, and the six-language system is used as starting point. Various methods are compared when the six-language system is adapted to a seven-language task. The proposed best method is further generalized to the nine-language task. The results show how we can efficiently adapt to new languages.

7.2.1 Data and Tasks

The data from the six languages are the same as those used in the six-language task discussed in the previous chapters. The seventh language used in this study is Vietnamese, because system performance on Vietnamese is representative in the results reported in previous chapters.

All the data from Vietnamese used in these experiments are from the training set and development set 2 of the eleven-language task described in Chapter 3. Compared with the systems described in the previous chapter, all the systems presented here use only approximately 60% of training data (in terms of hours) for Vietnamese (since the data in development set 1 are not used).

Besides Vietnamese, the other two languages used to perform the generalization task are Farsi and Tamil since they are used in the NIST'94 and NIST'95 evaluations. Again, only the data from the training set and development set 2 are used for these two languages.

7.2.2 The Approaches

Four approaches are studied. Three of them are based on the optimization method proposed in this thesis. In order to add a new language to the task, we need to estimate the LID models for this language in the score generators and retrain the neural networks (final classifier).

1. Baseline approach.

   The model estimation in this approach is the same as the implementation of our baseline system described in Chapter 4, except that less Vietnamese data are used. In our implementation, the models are estimated on the training set in the conventional way. The scores for Vietnamese to train the neural network are duplicated (slightly modified based on the corresponding variances) in order to balance the training patterns for each output node in the final neural-network classifier.

2. Optimization-based model estimation.

   In this approach, the models for the new language are initialized by averaging the existing models for all the other languages in the task. Then these initialized models are optimized using the training data for the new language with the optimization method described in Chapter 5. Again the training patterns for the neural network are balanced by duplication.

3. Interpolation.

   In this approach, the models for the new language are obtained by averaging the models obtained in the above two approaches.

4. Combined approach.

   This approach combines the above three approaches. All the training data for the new language are divided into three sets: training set, development set 1 and development set 2 (the ratio among these three sets are 6:3:1). The training set is used to estimate the model using the interpolated methods (Approach 3). The resulting models are then optimized using the development set 1 based on the cross-validation performance on development set 2. All the data are used in the training of the final classifier. The training patterns are also balanced by duplication.

All these four approaches are evaluated using the seven-language task (with English, German, Hindi, Japanese, Mandarin and Spanish as the existing languages in the task and Vietnamese as the new language). The systems are evaluated on data for these seven languages using the NIST'94 and NIST'95 test sets (whole utterance part). The results (correct rate) are reported in Table 7.2.

Table 7.2: Comparison of approaches to new language adaptation

Approach	Baseline	Optimization	Interpolation	Combination
NIST'94	89.0%	89.8%	91.3%	92.9%
NIST'95	88.6%	89.3%	91.4%	92.9%

We see that the combined approach outperforms the other approaches. In order to test the generalization ability of the proposed best method, the combined approach is extended to a system for the nine-language task (with Farsi, Tamil and Vietnamese as the new languages). The resulting system is evaluated on the NIST'94 and NIST'95 test data. The results are given in Table 7.3. For comparison, the results achieved by the final system described in Chapter 6 on these two test sets are also listed (labeled as *Conventional*) in the table.

Table 7.3: LID Results on new language adaptation: Nine-Language Task

Data	NIST94	NIST95
Conventional	92.5%	90.0%
Combined	91.9%	90.0%

7.2.3 Discussion

From the results of our comparative experiments, we see that each of the three proposed methods can outperform the baseline approach; approximately 35% error reduction is achieved by the combined methods.

The optimization procedure outperforms the baseline system by not over-fitting the model parameters. In the baseline system, the training patterns for the final classifier are derived from the data used to calculate the model parameters. There is thus a big difference between the testing and training patterns for the neural network (the training patterns are less noisy since they were used to calculate the parameters). The major drawback of the optimization procedure is that the initialization may not be appropriate. The interpolated method is designed to overcome the shortcomings of the first two approaches. Smoothing is a very important technique in model estimation. The most difficult aspect of this approach is how to select the best smoothing factor. In our experiment, we simply averaged the two models.

The combined method is proposed to further refine the interpolated approaches. The resulting models from the interpolated method are further optimized by the optimization procedure. The division of the training data ensures that there are new data for the training of each component of the LID system; this minimizes the possible mismatch between training patterns and testing patterns in each part. The generalization of this method to the nine-language task gives encouraging results. With 40% less training data, the system still achieves comparable results on the two NIST evaluation data sets. This demonstrates the feasibility of the proposed method for new-language adaptation at the expense of more training cycles.

Chapter 8

Conclusion

Encouraging language identification results have been achieved with the approach based on language-dependent phone recognition presented in this dissertation. The improved baseline system compared favorably with previously reported results on the same six-, nine- and eleven-language tasks.

8.1 Summary

Our evaluation of the different LID models on the same baseline system 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 language model plays the most important role in our system.

To address the contradictory requirements of detailed modeling and the availability of data, we proposed the backward bigram language model as an addition to the conventional forward bigram language model. Without drastically increasing the amount of training data required, it improves the level of detail in the language model. With the introduction of this model, we achieved approximately 20% error reduction on the whole utterance part on both tasks, and 10% error reduction on the ten-second long utterances, compared with the baseline system.

The introduction of the new duration model did not yield a great improvement. Although it helps the system slightly, effective modeling of the prosodic information is still an interesting issue for future research. The comparison of the neural-net classifier and the linear classifier showed that non-linear combination of these information sources is useful for language classification. Using the neural-net results in more than 20% error reduction on the whole utterance part and 15% error reduction on the ten-second part of both tasks compared with the linear classifier.

The proposed optimization method demonstrates the promise and feasibility of improving LID accuracy by increasing the discrimination of the language models using an automatic optimization procedure. With this back-propagation based optimization procedure, the detrimental effect of 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 whole utterance tasks and 15% error reduction on the ten-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 whole-utterance tasks and more than 20% error reduction on the ten-second utterance tasks. On the eleven-language task, more than 40% error reduction on the whole-utterance tasks and more than 25% error reduction on the ten-second utterance tasks were achieved, which shows that our methods for improving the system performance is generalizable to a larger task.

We achieved more improvement on the whole utterance tasks than on the ten-second utterance tasks. One major reason is that all our LID models are statistical models which are based on the decoded phone strings (the acoustic information is not exploited directly). On average the whole 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.

The preliminary studies on the processing of conversational speech and adaptation to new languages show that system performance can be improved by the proposed methods: more than 35% error reduction was achieved on both tasks.

8.2 Future Work

How to enhance the acoustic modeling is the key point to the identification of very short input utterances (and may also be highly beneficial for longer utterances). This is still an unsolved problem and was not addressed in this thesis work. With very short input utterances, systems highly reliant on phonotactic constraints can not get a reliable estimation of the underlying phone (or other subword units) sequential information. When high-level information is not available, acoustic information will be a predominant discriminant. The commonly used methods for acoustic modeling for LID research today are obtained directly from speech recognition techniques designed mainly for single languages; these techniques may not be appropriate for LID purposes, since an LID system will be exposed to multi-language signals. New acoustic optimization methods designed for LID purposes may further improve system performance by increasing the accuracy of acoustic modeling.

The robust modeling of prosodic information will be another interesting issue. How to differentiate the speaker-specific and language-specific information is critical to the success of an LID system. Current techniques focus on the variations within segments; normalization of prosodic information based on speech rate may be an efficient way to minimize speaker variabilities. Also, prosodic information is strongly correlated for contiguous segments; development of an intra-segment model of prosodic information can further improve the system performance. In this thesis work, only duration information is exploited; how to incorporate pitch and stress information into current LID model sets will also be part of further work.

From the confusion matrices of the system, we found that some languages were poorly identified; explicitly adding knowledge on these languages may further improve the system performance. For example, adding typical language-dependent phone strings (with variable length) into model sets (or directly employing keyword-spotting techniques) will increase the discrimination between languages. Directly modeling the language-dependent phones at acoustic level may be another way to enhance the discrimination.

For real-world applications, handling channel variation and providing a confidence measurement (for rejection) in the system are also very important issues.

As shown in the closed-set test, system performance decreased drastically because of the utterances from background languages. This suggests that it is impossible to build a robust system by just using maximum likelihood criterion. Other measurements such as mutual-information entropy, MAP or other criteria may be used for confidence estimation; these measurements can be employed in different system components and will be combined to provide a joint decision.

For two new pursuits—processing of conversational speech and adaptation to new languages—only preliminary studies were conducted for this thesis. For processing of conversational speech, although the proposed methods effectively decreased the system error rate, the overall performance is still far from what can be got on monologue speech. As discussed in Chapter 7, there are many differences between monologue and conversational speech. The improvement we achieved is mainly due to the handling of long silence segments and correcting of some obvious errors by the phone recognizers. There are other phenomena such as filled pauses, false starts and repetitions which have not been treated properly. By eliminating these differences, similar performance could be expected on these two kinds of inputs even if the models are still trained on monologue data. Explicit modeling of filled pauses may be another interesting way to improve system performance since they occur very often in conversations and are generally language-dependent. For adaptation to new languages, our preliminary experiments focus only on the adaptation of language models. Adaptation of other models can also be beneficial. It is also reasonable to expect that systems based on acoustic and prosodic information sources will excel in this department, since such systems do not require phonetically labeled data for training.

Bibliography

[ADB94] O. Andersen, P. Dalsgaard, and W. Barry. On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four european languages. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 121–124, April 1994.

[Ata74] B.S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55:1304–1312, 1974.

[BABC94] K. Berkling, T. Arai, E. Barnard, and R.A. Cole. Analysis of phoneme-based features for language identification. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 289–292, April 1994.

[BC89] E. Barnard and R.A. Cole. A neural-net training program based on conjugate-gradient optimization. Technical Report CSE 89-014, Computer Science Department, Oregon Graduate Institute, 1989.

[CI82] D. Cimarusti and R.B. Ives. Development of an automatic identification system of spoken languages: Phase 1. In *Proceedings 1982 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1661–1664, May 1982.

[DM80] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic,Speech and Signal Processing*, 28:357–366, August 1980.

[DW83] S. Dowdy and S. Wearden. *Statistics For Research.* John Wiley & Sons, Inc., 1983.

[Foi86] J.T. Foil. Language identification using noisy speech. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 861–864, April 1986.

[Fuk90] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, Inc., second edition, 1990.

[GHM92] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research development. In *Proceedings 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, March 1992.

[GMW89] F.J. Goodman, A.F. Martin, and R.E. Wohlford. Improved automatic language identification in noisy speech. In *International Conference of the American Society for Signal Processing*, pages 528–531, May 1989.

[HMH93] H. Hermansky, N. Morgan, and H.G. Hirsh. Recognition of speech in additive and convolutional noise based on rasta spectral processing. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 83–86, April 1993.

[HN77] A.S. House and E.P. Neuberg. Toward automatic identification of the language of an utterance: Priliminary methodological considerations. *Journal of the Acoustical Society of America*, 62(3):708–713, 1977.

[HTG95] S. Hutchins and A. Thyme-Gobbel. The role of prosody in language identification. In *the Fifteenth Annual Speech Research Symposium XV*, pages 76–83, June 1995.

[HZ93] T.J. Hazen and V.W. Zue. Automatic language identification using a segment-based approach. In *Proceedings Eurospeech 93*, pages 1303–1306, September 1993.

[HZ94] T.J. Hazen and V.W. Zue. Recent improvements in an approach to segment-based automatic language identification. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1883–1886, October 1994.

[Jel90] F. Jelinek. Self-organized language modeling for speech recognition. In K.F. Lee and A. Waibel, editors, *Readings in speech recognition*, pages 450–506. Morgan Kaufmann, 1990.

[KH94] S. Kadambe and J.L. Hieronymus. Spontaneous speech language identification with a knowledge of linguistics. In *Proceedings of International Conference on Spoken Language Processing*, pages 1879–1882, October 1994.

[KH95] S. Kadambe and J.L. Hieronymus. Language identification with phonological and lexical models. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3507–3510, May 1995.

[Lad93] P. Ladefoged. *A Course in Phonetics.* Harcourt Brace Jovanovich, third edition, 1993.

[LD74] R.G. Leonard and G.R. Doddington. Automatic language identification. Technical Report RADC-TR-74-200, Air Force Rome Air Development Center, August, 1974.

[LE80] K.P. Li and T.J. Edwards. Statistical models for automatic language identification. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 80*, pages 884–887, April 1980.

[LG93] L.F. Lamel and J.S. Gauvain. Identifying non-linguistic speech features. In *Proceedings Eurospeech 93*, pages 23–30, September 1993.

[LG94] L.F. Lamel and J.S. Gauvain. Language identification using phone-based acoustic likelihoods. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 293–296, April 1994.

[Li94] K.P. Li. Automatic language identification using syllabic spectral features. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 297–300, April 1994.

[Li95] K.P. Li. Experimental improvements of a language id system. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3515–3518, May 1995.

[MBA$^+$93] Y.K. Muthusamy, K.M. Berkling, T. Arai, R.A. Cole, and E. Barnard. A comparison of approaches to automatic language identification. In *Proceedings Eurospeech 93*, pages 1307–1310, September 1993.

[MBC94] Y.K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, October 1994.

[MBMB95] W.J. Mistretta, M. Birnbaum, D.P. Morgan, and K. Brown. Phoneme-based language identification incorporating segmental features. In *the Fifteenth Annual Speech Research Symposium XV*, pages 72–75, June 1995.

[MC92] Y.K. Muthusamy and R.A. Cole. A segment-based automatic language identification system. In J.E.Moody, S.J.Hanson, and R.P.Lippmann, editors, *Advances in Neural Information Processing Systems*, pages 241–248. Morgan Kaufmann, 1992.

[MCO92] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika. The ogi multi-language telephone speech corpus. In *Proceedings International Conference on Spoken Language Processing 92*, pages 895–897, October 1992.

[MJC94] Y.K. Muthusamy, N. Jain, and R.A. Cole. Perceptual benchmarks for automatic language identification. In *International Conference on Speech and Signal Processing*, pages 333–336, April 1994.

[Mut93] Y.K. Muthusamy. *A Segmental Approach to Automatic Language Identification.* PhD thesis, Oregon Graduate Institute, July 1993.

[NUS92] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *Proceedings International Conference on Spoken Language Processing 92*, pages 1011–1014, October 1992.

[PB92] D.B. Paul and J.M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings International Conference on Spoken Language Processing 92*, pages 899–902, October 1992.

[PC95] E.S. Parris and M.J. Carey. Language identification using phoneme recognition phonotactic language modeling. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3503–3506, May 1995.

[RMM91] L. Riek, W. Mistretta, and D. Morgan. Experiments in language identification. Technical Report SPCOT-91-002, Lockheed Sanders Inc., 1991.

[RR94] P. Ramesh and D.B. Roe. Language identification with embedded word models. In *Proceedings of International Conference on Spoken Language Processing*, pages 1879–1882, October 1994.

[RSN94] A.A. Reyes, T. Seino, and S. Nakagawa. Three language identification methods based on hmms. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1895–1898, October 1994.

[SAG91] M. Savic, E. Acosta, and S.K. Gupta. An automatic language identification system. In *International Conference of the American Society of Signal Processing*, pages 817–820, 1991.

[Sug91] M. Sugiyama. Automatic language recognition using acoustic features. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 813–816, may 1991.

[TCP94] R.C.F. Tucker, M.J. Carey, and E.S. Parris. Automatic language identification using sub-word models. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 301–304, April 1994.

[UN90] Y. Ueda and S. Nakagawa. Prediction for phoneme/syllable/word category and identification of language using hmm. In *Proceedings International Conference on Spoken Language Processing 90*, pages 1209–1212, November 1990.

[WLO$^+$95] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, and S.J. Young. The 1994 htk large vocabulary speech recognition system. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 73–76, May 1995.

[YB95a] Y. Yan and E. Barnard. Recent improvements to a phonotactic approach to language identification. In *the Fifteenth Annual Speech Research Symposium XV*, pages 212–219, June 1995.

[YB95b] Y. Yan and E. Barnard. An approach to automatic language identification based on language-dependent phone recognition. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3511–3514, May 1995.

[YB95c] Y. Yan and E. Barnard. An approach to automatic language identification with enhanced language model. In *Eurospeech Proceedings*, pages 1351–1354, September 1995.

[YB95d] Y. Yan and E. Barnard. Neural networks and linear classifiers automatic language identification. In *International Conference on Neural Networks and Signal Processing (ICNNSP95), To be published*, December 1995.

[YWB94] S.J. Young, P.C. Woodland, and W.J. Byrne. Spontaneous speech recognition for the credit card corpus using the htk toolkit. *IEEE Transactions on speech and audio processing*, 2:615–621, October 1994.

[Zis93] M.A. Zissman. Automatic language identification using gaussian mixtures and hidden markov models. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 399–402, April 1993.

[Zis95] M.A. Zissman. Language identification using phoneme recognition phonotactic language modeling. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3503–3506, May 1995.

[ZS94] M.A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modelling. In *Proceedings 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 305–308, April 1994.

Appendix A

Models Used in the Final System

Three sets of models are used; details can be found in Chapter 4 and 5.

- Forward-language Model

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

- Backward-language Model

$$P_{LB} = \prod_{i=1}^{T}(\alpha P_{LB}(O_i|O_{i+1}) + \beta P(O_i))$$

- Duration Model

$$P_D = \prod_{i=1}^{T}((\mathit{1} - \alpha)P_D(D[O_i]|O_i, O_{i-1} \in S) + \alpha P(D[O_i]|O_i))$$

Biographical Note

Yonghong Yan was born on March 16, 1967, in Wuxi City, Jiangsu Prov., P.R.China. He graduated from Jiangsu Changzhou High School with an Outstanding Student honor (July, 1985), and was admitted to Tsinghua University (Beijing, P.R.China) with the exemption from the National College Entrance Examination. He was honored as an Outstanding Student in the Electronics Engineering Department (1985-1987). In 1990, he received his Bachelor of Engineering in E.E..

During 1990-1992, he worked in Beijing Xinghe Institute of Intelligent Computer, which was one of the most successful speech labs in China at that time. He was the head of the speech recognition group in the Institute during 1991-1992. His research interests include signal processing, speech recognition, perceptual phenomena, language identification and real-time system implementation. He has been working in speech processing since 1986, and involved in the development of various speech systems.

Publication list:

- Y. Yan & E. Barnard. An approach to automatic language identification based on language-dependent phone recognition. In *Proceedings 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Pages V-3511 – V-3514, Detroit, Michigan, May, 1995.

- Y. Yan & E. Barnard. Recent improvements to a phonotactic approach to language identification. In *the Fifteenth Annual Speech Research Symposium XV*, Pages 212-219, Baltimore, Maryland, June, 1995.

- Y. Yan & E. Barnard. An approach to automatic language identification with enhanced language model. In *Eurospeech Proceedings*, Pages 1351-1354, Madrid, Spain, September, 1995.

- Y. Yan & E. Barnard. Neural Networks and Linear Classifiers Automatic Language Identification. Accept for *International Conference on Neural Networks and Signal Processing*, Nanjing, P.R. China, December, 1995.

- P. Vermeulen, E. Barnard, Y. Yan, M. Fanty & R.A. Cole. A Comparison of HMM and neural network approaches to real world telephone speech applications. Accept for *International Conference on Neural Networks and Signal Processing*, Nanjing, P.R. China, December, 1995.

- Y. Yan, E. Barnard & R.A. Cole. Development of An Approach to Automatic Language Identification based on Phone Recognition. Accept for Journal *Computer, Speech & Language.*