Automatic Language Identification Using a Segment-Based Approach

by

Timothy J. Hazen

S.B., Massachusetts Institute of Technology, 1991

Submitted to
the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Master of Science

at the

Massachusetts Institute of Technology

September, 1993

Signature of Author ....
Department of Electrical Engineering and Computer Science
May 9, 1995

Certified by ....
Victor W. Zue
Principal Research Scientist
Department of Electrical Engineering and Computer Science

Accepted by ....
Frederic R. Morgenthaler
Chair, Department Committee on Graduate Students

Automatic Language Identification Using a Segment-Based Approach

by

Timothy J. Hazen

Submitted to the Department of Electrical Engineering and Computer Science in August, 1993 in partial fulfillment of the requirements for the Degree of Master of Science

Abstract

Automatic Language Identification (ALI) is the problem of automatically identifying the language of an utterance through the use of a computer. In 1977, House and Neuburg proposed an approach to ALI which focused on the phonotactic constraints of different languages. Their work suggested that simple language models could be used effectively for language identification if an accurate phonetic representation of an utterance could be obtained from the acoustic signal. Our research utilizes House and Neuburg's ideas as the starting point for a new segment-based approach to ALI.

To develop a solid theoretical basis for the design of an ALI system, a formal probabilistic framework has been developed. This framework uses House and Neuburg's ideas as its foundation but also utilizes additional information that may be useful for ALI. Specifically, phonotactic, acoustic and prosodic information are all incorporated into the framework which provides the structure for the segment-based system.

To investigate the capabilities of the new segment-based approach, the system was trained and tested using the OGI Multi-Language Telephone Speech Corpus, which consists of utterances in 10 different languages. The entire system was able to identify the language of a test utterance 48.6% of the time. To investigate the system's performance in more detail, the entire system, as well as each component of the system, was evaluated as various test conditions were altered. Overall, the analyses of the system confirmed that the phonotactic constraints of languages can be used effectively for ALI. However, it was also discovered that additional information, such as prosodic and acoustic information, can also be useful to supplement the phonotactic information.

Thesis Supervisor: Victor W. Zue
Title: Principal Research Scientist

Acknowledgments

I wish to express my deepest gratitude to my thesis advisor, Victor Zue, for the support, encouragement, and advice he has extended to me over the last two and a half years. I will be ever grateful for the opportunity to learn from and work with Victor. His pleasant demeanor and humor has allowed the time I've spent in his group to be both productive and fun.

I also wish to thank the members of the Spoken Language Group for all of the support and friendship they have provided me. In particular I would like to thank:

Mike Phillips for his assistance in adapting SUMMIT to fit my needs.

Jim Glass for his efforts in developing the mixture Gaussian code that was used in this thesis.

Stephanie Seneff and Jim Glass for reading a draft of my thesis and providing helpful suggestions for the final version.

Mike McCandless for his help on language modeling.

David Goodine, Joe Polifroni, and Christine Pao for keeping our systems up and running.

Vicky Palay for everything she's done to keep our group running smoothly.

Helen Meng for her encouragement and for putting up with my constant ramblings about my work.

Bill Goldenthal and Jane Chang for providing unending amusement whenever I needed a break from my work.

Lee Hetherington, Rob Kassel, Nancy Daly, and David Goddeau for all the help they've provided and questions they've answered.

Special thanks also go out to Yeshwant Muthusamy and Ron Cole for collecting a wonderful corpus for ALI research and providing it to our group as early as was possible.

Finally, I wish to thank my parents for all their love and support.

This research was supported by ARPA under Contract N0014-89-J-1332 monitored through the Office of Naval Research and by a grant from Texas Instruments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Automatic Language Identification (ALI) is the problem of automatically identifying the language of a spoken utterance through the use of a computer. Although research of the ALI problem began over 20 years ago, until recently there have only been a handful of published studies conducted on the topic. Early interest in the ALI problem originated within the intelligence community where automated language identification could provide obvious benefits. More recently, with increased activities in the development of multi-lingual speech recognition/understanding systems, interest in ALI has spread into the academic and industrial communities as well. Applications such as machine translation and multi-lingual information retrieval could benefit greatly if effective methods for identifying the language a person is speaking can be developed.

Figure 1.1 shows how a language identifier would fit into a multi-lingual information retrieval system. For this system the job of the language identifier is to determine what language is being used in the incoming utterance so that the utterance can be passed to the proper speech recognition/understanding system. Ideally the language identifier should achieve a high accuracy rate in identifying the language of spoken utterances while also being computationally efficient. However, in reality one must consider the tradeoff between accuracy and efficiency.

Upon initial examination of the language identification problem, one may note that each language of the world can be distinguished from any other language by its own unique vocabulary. However, utilizing knowledge about the unique vocabulary of each language would also require a knowledge of the syntactic and semantic rules which govern the concatenation of words into spoken utterances. Clearly it would be possible to develop a nearly flawless ALI system if this information could be successfully incorporated into a system. By example, this approach to ALI could be handled by simply developing a speech recognition system for all possible languages. The lan-

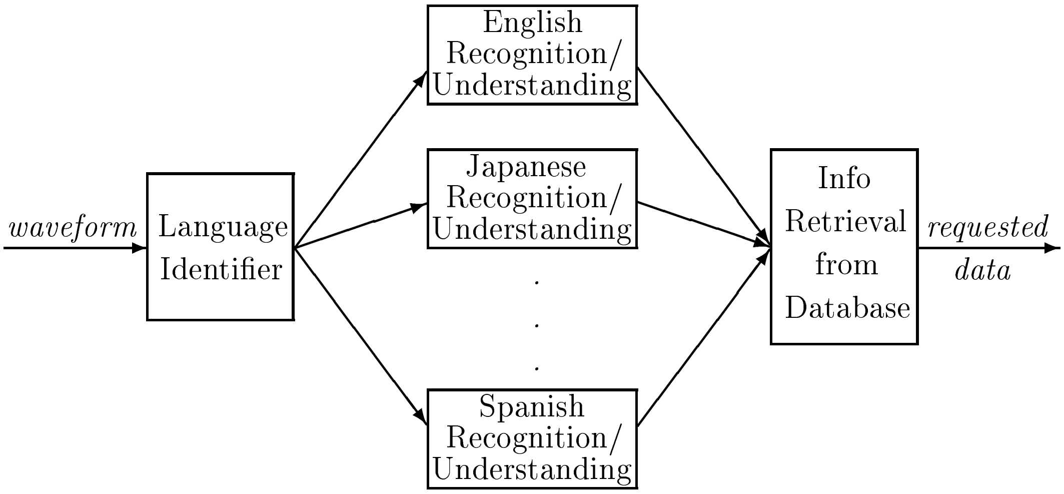

Figure 1.1: A multi-lingual system using a language identifier

guage of an utterance would be determined when the recognizer trained for the correct language is able to produce a viable string of words to match the waveform, while the recognizers for other languages are unable to decipher the input. However, there are two main reasons why this type of approach may be impractical. First, extensive expert knowledge of multiple languages may require a tremendous effort to collect, organize, and incorporate into an ALI system. Second, even if extensive expert knowledge is available and can be incorporated into a system, it may be computationally impractical to use *all* of this knowledge to identify the language that is being spoken. Thus, the goal of ALI research to date has generally been to develop dependable language identification methods which do not rely upon higher-level knowledge of the languages involved. Additionally, many past ALI studies have concentrated on utilizing *only* the information that is *directly* available from the waveform (i.e., acoustic features).

It appears plausible that accurate ALI may be achieved utilizing only the information that is available from the waveform of a spoken utterance. It has been observed that humans often have the ability to identify the language of a spoken utterance even when they have no working knowledge of the vocabulary or syntax of that language [28]. As will be discussed in Chapter 2, an investigation into the properties of different languages reveals that languages often differ in their phonological and prosodic characteristics. These characteristics are evident in the waveform of a spoken utterance. It is the differences in these characteristics which has motivated all of the ALI approaches to date, including the research presented in this thesis.

1.2 Previous Work

The most common general approach to the ALI problem has been the use of frame-based statistical methods. These methods use acoustic models to identify the language of a spoken utterance based on the frame by frame statistics of the utterance's acoustic features. The studies by Cimarusti and Ives [1], Ives [13], Foil [6], Goodman et al. [10], Sugiyama [34], Savic et al. [32] and Zissman [36] are similar in that each used a frame-based language identification algorithm which was trained on acoustic features of the speech signal in an unsupervised fashion. Thus, none of these studies used any prior knowledge of the underlying phonetic or prosodic structure of their data. While the specifics of the classification algorithms are different in each case, each algorithm was designed to identify the language of an utterance based only on the statistics of acoustic features. None of these approaches attempts to model the speech as a sequence of linguistic events.

The earliest published research in ALI in this country was performed by Leonard and Doddington [18, 19, 20, 21]. They developed an approach where language identification was performed by identifying sound segments or sequences which are particular or common to specific languages. Once a set of useful sound segments was proposed, language identification was performed by examining the probability distribution of the selected sound segments within a speech utterance. This approach was based on the assumption that certain linguistic events occur more frequently in particular languages and the observed statistics of these events can provide for accurate language identification.

A similar approach to ALI was proposed by House and Neuburg [11]. Like Leonard and Doddington, they believed language identification could be performed by observing the statistics of the linguistic events present in a speech utterance. More specifically, they believed that languages could be identified based upon the sequential constraints of their phonetic elements. Based on this belief they proposed a two step approach to ALI. The first step was to transform an utterance into a string of phonetic elements. The second step was to identify the language of the utterance by examining the statistics of the phonetic sequence. However, they believed that the extraction of a *detailed* phonetic sequence from a spoken utterance of an unknown language could not be performed with sufficient reliability and, in fact, might not even be necessary. Instead, they proposed an approach where the speech input was transformed into a sequence of broad phonetic classes. They believed the automatic extraction of the underlying string of broad phonetic classes of a spoken utterance could be performed with high reliability, though they did not confirm this hypothesis on actual speech data. However, in a feasibility study, they did confirm their belief that the statistics of sequences of broad phonetic classes would be sufficient for reliable language identification given a long enough phonetic sequence. They showed this empirically by transcribing texts from eight different languages into strings of five broad phonetic classes and evaluating bigram and trigram models applied to these transcribed texts.

The results presented by House and Neuburg offer the hope that very simple phonetic language models can be powerful tools for language identification. While their work solidly showed that simple phonetic language models work exceptionally well when the underlying string of broad phonetic classes for an utterance is known exactly, they did not prove that these language models could be robust when the string of phonetic classes contained errors. However, a few studies that utilize House and Neuburg's basic premise have been conducted.

The work of Li and Edwards [23] was the first attempt following the general framework proposed by House and Neuburg to be tested on actual speech data. They designed a frame-based classifier which labeled each frame of an utterance with a broad phonetic class. Using a post-processing smoothing algorithm, they transformed the frame-based sequence of phonetic labels into a sequence of segments labeled with broad phonetic classes. The language identification was then performed using various finite state statistical models on the sequences of broad phonetic classes. Unfortunately, their study demonstrated that House and Neuburg's approach was effective but not infallible. Their results showed that the use of an imperfect phonetic recognizer for determining the string of broad phonetic classes clearly hurt the ability of the language models to perform highly accurate language identification.

A study by Muthusamy and Cole [27, 28] also utilized the idea of transforming the input speech into a sequence of broad phonetic classes. However, they did not limit the language identification process to simply building language models for the phonetic class sequence. Instead, they devised an approach where various phonetic and prosodic features were extracted from the segments of the phonetically labeled utterance. A neural network which was trained using these features was then used to perform the language identification.

Lamel and Gauvain [16] used an approach where a phonetic recognition system was trained separately for each language. The training produced language dependent phone and language models for each language. The language of a test utterance was then determined by applying each language dependent phonetic recognizer to the utterance and choosing the specific recognizer which produced the highest normalized likelihood score (i.e., the recognizer which was able to produce the closest match between the waveform and its own language specific models). Lamel and Gauvain only tested their approach on the two language set of English and French. For large language sets this approach could become computationally burdensome.

It is very difficult to determine which of the above approaches to the ALI problem are the most effective. For the most part, each of the studies mentioned above utilized a different speech corpus. These corpora varied over many different conditions includ-

Authors of Study	Year	Number of Languages Used	Avg. Length of Test Utterances	Reported Accuracy
Leonard	1980	5	60s	72%
Li and Edwards	1980	5	120s	78%
Cimarusti and Ives	1982	5	?	84%
Ives	1986	5	?	92%
Sugiyama	1991	20	64s	80%
Muthusamy and Cole	1992	10	13.4s	48%
Zissman	1993	10	13.4s	46%
Lamel and Gauvain	1993	2	4s	100%

Table 1.1: Summary of previous published results

ing their language sets, bandwidths, channel characteristics, vocabulary constraints, and test utterance lengths. Without a common set of test conditions, a meaningful comparison of the results reported in the different studies is not possible. Nevertheless, a brief summary of the results that have been published is shown in Table 1.1. It should be mentioned that the Muthusamy and Cole system and the Zissman system were both tested on the OGI Multi-Language Telephone Speech Corpus [29]. This is the same corpus that was used for the experiments that are presented in this thesis.

1.3 Thesis Overview

The ultimate goal of ALI research is to develop language identification methods which are reliable, computationally efficient, and easily portable to new language sets. However, the scope of this thesis is limited to research towards the development of a reliable ALI approach which does not require higher level knowledge of the languages it is attempting to identify. The research presented in this thesis does not consider the issues of computational efficiency or portability to new languages. In its investigation of the ALI problem, the basic goals of this thesis can be summarized as:

1. Present a formal probabilistic framework describing the ALI problem.

2. Present a new segment-based approach to the ALI problem.

3. Analyze and understand the various modeling decisions, assumptions, and test conditions which affect the performance of the system.

As the starting point for the development of a new ALI design, a formal probabilistic framework of the ALI problem has been derived. Unlike the automatic speech recognition problem, no formal probabilistic framework describing the ALI problem has been presented in any of the existing papers on the subject. Such a framework is presented in Chapter 2. It utilizes House and Neuburg's ideas as a foundation and provides the structure for the ALI design which is described in this thesis.

Utilizing the probabilistic framework, a new segment-based approach to the ALI problem has been developed. Like Muthusamy and Cole's system, the new design retains the basic ideas of House and Neuburg while also allowing for additional information to be used in the language identification process. The basic architecture of the new system is shown in Figure 1.2. In this diagram $\vec{\mathbf{a}}$ and $\vec{\mathbf{f}}$ represent the acoustic and fundamental frequency information that is extracted from the waveform, C represents the string of phones or broad phonetic classes that the phonetic recognizer extracts from the acoustic information, and S represents the segmentation of the waveform which matches the phonetic string C. In this design the language identifier may use *any* information that is available from the acoustic feature vectors, fundamental frequency contour, phonetic sequence or segmentation. A detailed description of each of the components in this system is provided in Chapter 3. An analysis and evaluation of the performance of the new system is presented in Chapters 4 and 5.

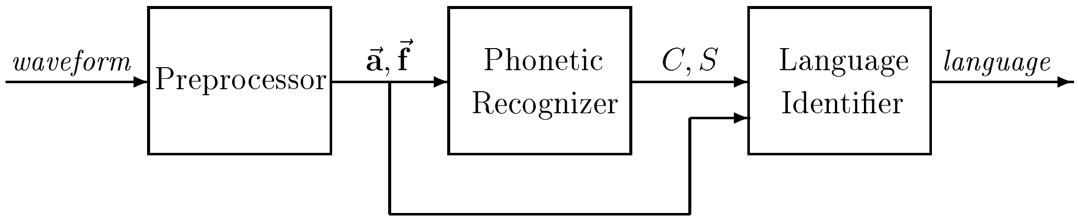

Figure 1.2: Proposed ALI Design

Chapter 2

Theory

2.1 Discriminative Information for ALI

2.1.1 Overview

The design of a successful ALI system must begin with an understanding of the characteristics of spoken language which are most useful for the purpose of language identification. An ALI system needs to exploit the primary differences which exist among languages while still being robust in the face of speaker, channel and vocabulary variability. However, the system also needs to be computationally efficient. Thus, it is desirable to discover language discriminating characteristics which are relatively easy to extract from the acoustic signal, do not require complex methodologies to model, and are relatively free of noise from speaker, channel and vocabulary dependencies.

As discussed in Chapter 1, it may be possible to develop an ALI system which can accurately identify languages based only on information that is directly available from the waveform of a spoken utterance. The information that is available in an utterance's waveform can be viewed as belonging to one of two groups, phonological information and prosodic information. The series of spoken sounds (or phones) which is present in the spoken utterance contains the phonological information. The fundamental frequency, intensity and duration variations that span across the spoken utterance contain the prosodic information. While the phonological and prosodic information available in the signal may represent some higher level information which is useful in determining the semantics of an utterance, knowledge of this higher level information may not be needed to identify the language of the utterance. It is hoped that adequate language identification can be performed using only the phonological and prosodic information of an utterance.

2.1.2 Phonological Information

The phonological properties of a spoken utterance can vary greatly from language to language. There are various phonological factors which help define the distinctiveness of a language. Some of these factors include the phone set, the phonotactic constraints and the acoustic realizations of particular phones within a language.[1]

Because each language uses only a small subset of phones from the set of all possible speech sounds which exist, variances can be observed across the phone sets of different languages [31]. Thus, a knowledge of the phones used in particular languages may be enough to help distinguish one language from another. Even if languages contain nearly identical phone sets, the languages may still be distinguishable by the probability distribution of the phones across each language. Thus, a phone that is commonly used in one language may be used rarely in another.

Different languages may also have different rules governing how sequences of phonemes may be constructed to form higher level linguistic elements such as syllables or words. These phonotactic constraints could cause certain phonetic sequences to be likely in some languages but unlikely in others. For example, Japanese has strict phonotactic constraints which generally prohibit consonants from following consonants. English, on the other hand, has looser constraints which allow for the possibility of multiple consonants in succession.

Significant differences may also exist in the acoustic realization of particular phones across different languages. These differences may be caused by cross language differences in the articulatory gestures used to produce the phone. For example, the phoneme /t/ can be realized by a large set of allophones. It can be realized with or without aspiration, with a dental or alveolar closure, and with lips rounded or unrounded. The use of each of these allophones varies across languages. Some differences in the acoustic realizations of particular phones across languages may occur because of the particular phonotactic constraints present within each language. The phonotactic constraints of different languages may cause certain coarticulation effects to be possible in one language and not possible in another. Thus, the differences that arise in the acoustic realizations of phones may be useful for distinguishing languages.

[1] For clarification, the difference between a *phoneme* and a *phone* should be stated. A phoneme is strictly a linguistic unit. A phone is a particular speech sound. A phone can viewed as the acoustic realization of a phoneme. Since higher level linguistic knowledge is not being used in the ALI design presented in this thesis, knowledge of the particular phonemes that exist in each language is not as important as knowledge of the particular phones that exist in each language. Thus, all references to phonetic elements, sequences, etc. that are made within this thesis refer to the phones within an utterance and not the phonemes.

2.1.3 Prosodic Information

The prosodic properties of languages can also vary greatly. Fundamental frequency (F0), duration and voice intensity are all important elements used within the prosodic structure of a spoken utterance. The manner in which these elements are incorporated into the prosodic structure of an utterance varies across languages. The differences across languages can often be observed in the realization of the prosodic features which determine the tones or stress contained throughout an utterance.

In tonal languages such as Chinese, the F0 contour and segment duration are used in determining the tone attached to a particular phone. Altering the tone for a particular phone can completely change the meaning of the word to which the phone belongs. Thus, in a tone language the F0 and phone duration patterns are strongly dependent on the types of tones used in that language and their relative probability distributions.

In languages that incorporate the concept of word stress, the intensity, duration, and F0 contour of a syllable are all correlated with the inherent stress being placed on that particular syllable [35]. Different languages use stress in different manners. For free stress languages, such as English, the stress pattern of words can vary between words with the same number of syllables. For fixed stress languages, such as Polish, the stress pattern is dependent only on the number of syllables present in each word. Thus, two words with the same number of syllables will always have the same stress pattern [31]. The exact manner in which F0, duration and intensity contribute to the stress of a syllable may also differ from language to language. For example, the timing of rises or falls of the F0 contour in relation to the placement of stressed syllables can vary. Some languages use a rising F0 at the beginning of a stressed syllable while others use a rising F0 at the end of a stressed syllable [35].

It has also been observed that some languages use the F0 contour to represent even higher level linguistic information. The F0 contour of the end of an utterances has been observed to differentiate between declarative statements and yes/no questions in languages such as English, French, Italian, and Japanese [35, 2]. In some languages such as English, declarative statements are characterized by a falling F0 contour at the end of an utterance while yes/no questions are characterized with a rising contour. However, other languages have been observed to contain just the opposite, a rising contour for declaratives and a falling contour for questions.

The effect of prepausal lengthening of vowels is another prosodic effect which has been observed to differ across languages. Lengthening of the final vowel in a sentence is a readily observable characteristic of spoken utterances in English, French, German and Italian. However, other languages such as Finnish, Estonian, and Japanese have been observed to contain little to no sentence-final lengthening of vowels [35].

2.2 Probabilistic Framework

2.2.1 Maximum *A Posteriori* Probability Approach

General Derivation

Before designing any system, it is desirable to develop a strong theoretical framework on which the design can be based. For this thesis the framework will be probabilistic in nature. To begin, let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ represent the language set of n different languages. When an utterance is presented to the ALI system, the system must use the acoustic information to decide which of the n languages in $\mathbf{L}$ was spoken.

Typically, the acoustic information of a spoken utterance is represented as a sequence of feature vectors where each individual vector represents the acoustic information for a particular time frame. For this derivation, it will be assumed that two specific types of information will be extracted from the waveform for each time frame; these are the wide-band spectral information and the voicing information. The wide-band spectral information is the most useful information for determining the underlying phonetic sequence of a spoken utterance. The voicing information, i.e. the F0 contour, is primarily used in describing the prosody of an utterance. Because of the separate natures of the two types of information, it is useful to represent them as two separate sequences of vectors. Therefore, let $\vec{\mathbf{a}} = \{\vec{\mathrm{a}}_1, \vec{\mathrm{a}}_2, \ldots, \vec{\mathrm{a}}_m\}$ be the sequence of m vectors which represent the wide-band spectral information of a spoken utterance and let $\vec{\mathbf{f}} = \{\vec{\mathrm{f}}_1, \vec{\mathrm{f}}_2, \ldots, \vec{\mathrm{f}}_m\}$ be the sequence of m vectors which represent the voicing information of a spoken utterance. To clarify the terminology used in this thesis, the wide-band spectral information contained in $\vec{\mathbf{a}}$ will be referred to as the acoustic information and the vectors contained in $\vec{\mathbf{a}}$ will be referred to as acoustic feature vectors. The information in $\vec{\mathbf{f}}$ will be referred to as the F0 information.

The probability that an utterance was spoken in language L_i, given the sequences $\vec{\mathbf{a}}$ and $\vec{\mathbf{f}}$, is represented by the expression $\Pr(L_i \mid \vec{\mathbf{a}}, \vec{\mathbf{f}})$. The maximum *a posteriori* probability (MAP) approach to the ALI problem is to choose the language which is most likely given the acoustic and F0 information. Mathematically this can be expressed as

$$\text{Choose } L_j \text{ such that } \Pr(L_j \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}) > \Pr(L_i \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}) \;\; \forall\, i \neq j. \tag{2.1}$$

Viewed as a maximization process the MAP approach can alternatively be expressed as

$$\arg\max_i \Pr(L_i \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.2}$$

The expression in (2.2) is the most general expression describing the ALI problem and should serve as the starting point for any probabilistic approach to ALI.

Incorporating Linguistic Information

Because each spoken utterance contains an underlying sequence of linguistic events, a probabilistic framework which incorporates linguistic information is appropriate. To incorporate this information into the framework, let $\mathbf{C}$ represent the set of all possible linguistic sequences which can represent a spoken utterance. Since phonetic elements are the most obvious choice for representing the linguistic sequence, it will be assumed that the sequences in $\mathbf{C}$ are represented with phonetic elements in the derivations that follow. Specifically, each unique phonetic sequence will be of the form $C = \{c_1, c_2, \ldots, c_p\}$ where each c is represented with a phonetic element. The set of elements that can be used in the phonetic sequence can be chosen to be as detailed as phones or as general as broad phonetic classes. However, the exact set of elements that are to be used in the design is not important for the derivation of the probabilistic framework. By incorporating the phonetic sequence into the framework, the expression in (2.2) becomes

$$\arg\max_i \sum_{\mathbf{C}} \Pr(L_i, C \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.3}$$

Proceeding from (2.3), there are two general categories of approaches which can be developed, frame-based and segment-based. In a frame-based approach, the probabilistic framework mandates that a phonetic element be associated with each single frame of the acoustic input. In a segment-based approach, the model assumes that sets of adjacent frames may underlyingly belong to the same phonetic element. Thus, in segment based approaches, only *one* phonetic element will be associated with each segment or set of related adjacent frames. These two different approaches are discussed separately below.

2.2.2 Frame-Based Approach

The main constraint in defining a frame-based approach is that each element of a phonetic sequence is mapped one-to-one with its corresponding acoustic frame. If $\vec{\mathbf{a}}$ contains m frames then all allowable phonetic sequences C must contain m elements.

In deriving any probabilistic approach, it is often useful to expand general probabilistic expressions such as (2.3) into multiple probabilistic terms which are simpler to model. To begin, (2.3) can be reworked as

$$\arg\max_i \sum_{\mathbf{C}} \frac{\Pr(L_i, C, \vec{\mathbf{a}}, \vec{\mathbf{f}})}{\Pr(\vec{\mathbf{a}}, \vec{\mathbf{f}})}. \tag{2.4}$$

Since the denominator in (2.4) is independent of i, it can be removed from the maximization process to yield

$$\arg\max_{i} \sum_{\mathbf{C}} \Pr(L_i, C, \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.5}$$

The expression can further be rewritten as

$$\arg\max_{i} \sum_{\mathbf{C}} \Pr(\vec{\mathbf{a}} \mid \vec{\mathbf{f}}, C, L_i) \Pr(\vec{\mathbf{f}} \mid C, L_i) \Pr(C \mid L_i) \Pr(L_i) \tag{2.6}$$

The transformation of (2.3) into (2.6) is useful because (2.6) is organized in such a fashion that the acoustic, F0, and phonetic sequence information can all be modeled separately. Despite the organization of the expression, it lacks a direct means of modeling the durations of the underlying phonetic elements. Because frame-based approaches do not incorporate the notion of segments, the information regarding the duration of the underlying phonetic elements is not explicitly available but rather is embedded within the sequence C.

It should be noted that the expression in (2.6) can be easily simplified to form the probabilistic description of a hidden Markov model (HMM) approach. The HMM approach has been widely used for many speech recognition related problems including ALI [32, 36]. The HMM approach can be formulated by applying the following assumptions:

1. $\vec{\mathbf{f}}$ is independent of $\vec{\mathbf{a}}$ and C.
2. The frames of $\vec{\mathbf{a}}$ are independent.
3. C is a Markovian sequence.

With these assumptions, the HMM approach can be represented by the expression

$$\arg\max_{i} \Pr(L_i) \Pr(\vec{\mathbf{f}} \mid L_i) \sum_{\mathbf{C}} \prod_{k=1}^{m} \Pr(\vec{\mathrm{a}}_k \mid c_k, L_i) \Pr(c_k \mid c_{k-1}, L_i). \tag{2.7}$$

2.2.3 Segment-Based Approach

For a segment-based approach, the concept of segmentation of the input speech must be incorporated into the probabilistic framework. To do this, let $\mathbf{S}$ represent the set of all possible segmentations of the input speech. In using a segment-based approach, the set of phonetic sequences that can belong to a particular segmentation is constrained by the assumption that there is a direct one-to-one mapping of phonetic elements to segments. To represent a particular segmentation containing p segments, let $S = \{s_1, s_2, \ldots, s_{p+1}\}$ where each s represents the location of a segment boundary. The only allowable set of phonetic sequences which can correspond to S are those with p phonetic elements. Thus, given S the phonetic sequence can be represented as $C = \{c_1, c_2, \ldots, c_p\}$. With these new considerations the maximization process in (2.3) can be expanded as

$$\arg\max_i \sum_{\mathbf{S}} \sum_{\mathbf{C}} \Pr(L_i, S, C \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.8}$$

This expression can be rewritten as

$$\arg\max_i \sum_{\mathbf{S}} \sum_{\mathbf{C}} \Pr(L_i \mid C, S, \vec{\mathbf{a}}, \vec{\mathbf{f}}) \Pr(C \mid S, \vec{\mathbf{a}}, \vec{\mathbf{f}}) \Pr(S \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.9}$$

Up to (2.9) no assumptions have been made, i.e., (2.9) is *exactly* equivalent to (2.2). However, with the tremendously large set of possible segmentations and phonetic sequences that could represent each utterance, it would be impractical to attempt to perform the summations in (2.9) over all S and all C. The required computation can be greatly reduced if only a subset of the possible segmentations and phonetic sequences are used in estimating the probabilities of each candidate language. These probabilities can potentially be estimated accurately using only the n-best phonetic hypotheses. To take this assumption even further, it may be feasible to assume that only the most likely segmentation and phonetic sequence needs to be found in the process of identifying the most likely language candidate. In this case, the expression in (2.9) can be reduced to

$$\arg\max_{i,\mathbf{S},\mathbf{C}} \Pr(L_i \mid C, S, \vec{\mathbf{a}}, \vec{\mathbf{f}}) \Pr(C \mid S, \vec{\mathbf{a}}, \vec{\mathbf{f}}) \Pr(S \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.10}$$

Additionally, it may be feasible to decouple the search for the most likely segmentation and phonetic sequence from the search for the most likely language. This would assume that the best segmentation and phonetic sequence can be found independent of the language of the utterance. The result of this assumption is that the maximization process in (2.10) can be separated into two steps. First, the most likely segmentation and phonetic sequence are found using

$$\arg\max_{\mathbf{S},\mathbf{C}} \Pr(C \mid S, \vec{\mathbf{a}}, \vec{\mathbf{f}}) \Pr(S \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.11}$$

Let the most likely segmentation and phonetic sequence be represented as $\hat{S}$ and $\hat{C}$. After $\hat{S}$ and $\hat{C}$ are found, the second step is to identify the most likely language using

$$\arg\max_i \Pr(L_i \mid \hat{C}, \hat{S}, \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.12}$$

As discussed previously with the frame-based approach, it may be useful to expand the general expression in (2.12) into multiple terms for the purpose of simplifying the modeling of the expression. To begin, it can be reworked as

$$\arg\max_i \frac{\Pr(L_i, \hat{C}, \hat{S}, \vec{\mathbf{a}}, \vec{\mathbf{f}})}{\Pr(\hat{C}, \hat{S}, \vec{\mathbf{a}}, \vec{\mathbf{f}})}. \tag{2.13}$$

In the maximization process, the denominator is constant across all i and can be removed leaving

$$\arg\max_i \Pr(L_i, \hat{C}, \hat{S}, \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{2.14}$$

This expression can be expanded into

$$\arg\max_i \Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i) \Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) \Pr(\hat{C} \mid L_i) \Pr(L_i). \tag{2.15}$$

The four probability expressions in (2.15) are considerably easier to model separately than the single probability expression in (2.12). Additionally, the expression is now organized in such a way that prosodic and phonetic information are contained in separate terms. In modeling, these terms become known as:

1. $\Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i) \rightarrow$ The acoustic model.
2. $\Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) \rightarrow$ The prosodic model.
3. $\Pr(\hat{C} \mid L_i) \rightarrow$ The language model.
4. $\Pr(L_i) \rightarrow$ The *a priori* language probability.

The prosodic model captures the differences that can occur in prosodic structures of different languages due to the stress or tone patterns created by variations in the phone durations and F0 contour. The phonetic information is divided into two separate models, the language model and the acoustic model. The language model will account for the probability distributions of the phonetic elements and the phonotactic constraints within each language. The acoustic model will account for the different acoustic realizations of the phonetic elements that may occur across languages. Aside from modeling concerns, this organization also provides a useful structure for evaluating the relative contributions towards language identification that phonotactic, prosodic, and acoustic information provide.

It should also be noted that while a maximum *a posteriori* probability approach was described in this derivation, the maximum likelihood approach can be achieved by simply ignoring the *a priori* language probability. In effect, this is identical to assuming all of the languages in the language set $\mathbf{L}$ are equally likely.

Chapter 3

System Design

3.1 System-Wide Decisions

3.1.1 Overview

Before making any detailed design and modeling decisions within the ALI system, a set of system-wide issues must first be resolved. These key issues can be summarized in the following questions:

- What is the goal of the ALI system?
- What type and amount of data will be used for training and testing?
- What criteria will be used to evaluate the system?

3.1.2 System Goals

From the outset, the goal of an ALI system must be defined. In particular, it must be decided whether the system will perform language verification or language recognition. Systems that perform language verification simply verify whether or not an utterance is spoken in one particular language. Systems that perform language recognition must identify the language of a spoken utterance from a set of language candidates.

If language recognition is the goal, it must also be decided whether the recognition will be performed with an open or closed set of languages. If a closed set of languages is used, the system will only be subjected to test utterances which are spoken in a language which is present in the system's training set. However, if an open set condition is used, the system may be presented with utterances which are spoken in languages which do not appear in its training set. In this case, the system needs to be able to *reject* any utterance which is spoken in a language that is not within it's training set.

The ALI system that is presented in this thesis is a language recognition system which operates on a closed set of languages. Thus, during testing, the system is not subjected to any languages which are not contained in the training set. We believe that it is necessary to first develop a concrete understanding of the issues involved in closed set language recognition before attacking the difficult issues involved in determining useful rejection criteria for the open set problem. Thus, this thesis only concentrates on the problem of reliable closed set language recognition.

3.1.3 Data Set

The data set that will be used must also be clearly defined. Because the discriminative information that is useful in language identification may vary from language set to language set, the set of languages should be clearly defined from the beginning. The amount of training data available in each language must also be carefully taken into account. The complexity of the models used in the system depends on the amount of available training data. Additionally, the system design should also consider the constraints placed on the vocabulary and context of the data, as well as the conditions under which the data set was recorded.

For this thesis, the ALI system is evaluated using the OGI Multi-Language Telephone Speech Corpus [29]. The corpus was collected at the Oregon Graduate Institute (OGI).[1] It contains utterances collected over the phone lines, at an 8 kHz sampling rate, from callers who were native speakers of one of ten different languages. These languages are English, German, French, Spanish, Farsi, Tamil, Vietnamese, Mandarin Chinese, Korean, and Japanese.[2] The utterances include fixed vocabulary utterances, topic-specific utterances, and unconstrained utterances. For each speaker up to ten utterances were collected. Four of the ten utterances contained a fixed vocabulary. Four others were text-independent but topic-specific. The final two were completely unconstrained. The prompts used to elicit the utterances from each speaker are described below along with the time allotted for the speaker's response to each prompt. It should be noted that a usable utterance was not always collected for each prompt.

[1]While the OGI corpus may eventually be used for many different topics in multi-lingual speech research, the corpus was originally collected by Yeshwant Muthusamy to aid his research in automatic language identification.

[2]As a reference, Appendix A contains a breakdown of the language families of each of the ten languages. Appendix B provides a table of the specific phones which are used in each language.

The fixed vocabulary utterances were elicited from each speaker with the following prompts in their native language:

1. What is your native language? (3 seconds)
2. What language do you speak most of the time? (3 seconds)
3. Please recite the seven days of the week. (8 seconds)
4. Please say the numbers zero through ten. (10 seconds)

The topic-specific utterances were the responses of each speaker to the following prompts:

1. Tell us something that you like about your hometown. (10 seconds)
2. Tell us about the climate of your hometown. (10 seconds)
3. Describe the room that you are calling from. (12 seconds)
4. Describe your most recent meal. (10 seconds)

The unconstrained utterances were collected by asking each speaker to speak freely about any topic of their choosing for one minute. Each unconstrained utterance was divided into two separate portions for the corpus; one with ten seconds of speech, the other with the remaining speech of the utterance.

The corpus is subdivided into three groups, a training set, a development test set and a final test set. The training set contains the utterances from fifty speakers in each language. The development test set contains twenty speakers for each language. The final test set contains twenty speakers from each language. For this thesis, the training set is used for training the ALI system and the development test is used for testing. The final test has been set aside for future work. Furthermore, only the topic-specific and unconstrained utterances are utilized. The fixed-vocabulary utterances are not used.

Excluding the fixed-vocabulary utterances, the training set contains 2715 utterances and the development test set contains 1120 utterances. The utterances are roughly evenly distributed amongst the ten languages. The number of utterances per speaker varies from 2 to 6. The male to female ratio of the speakers is roughly 7 to 3. Unfortunately some languages contain over 85 percent male speakers while others contain only 60 percent male speakers. It should also be noted that at the time of the experiments in this thesis the corpus had not yet been transcribed[3]. Without a full transcription of all of the utterances in the corpus, completely supervised training for phonetic recognition is not possible.

[3]As of the writing of this thesis, work is in progress at OGI to phonetically transcribe the utterances in the corpus.

3.1.4 System Evaluation

The measures of performance that are used to evaluate the system must also be defined. The most obvious measure of performance is the system's ability to reliably identify the language of a spoken utterance. Closely related to this is the system's ability to identify the language family of an utterance. However, aside from reliability in language identification, the system may also be evaluated based on other aspects such as its computational requirements, required training set size, and portability to different language sets. For this thesis, the design of the ALI system only considers the system's ability to perform reliable language identification.[4] Because reliable methods for ALI have not yet been developed, we believe it is best to develop insights into the primary problem of language recognition before other issues such as computation and portability are considered.

In evaluating the system's performance, two statistics are commonly used throughout this thesis. These statistics are the language identification accuracy and the rank order statistic. The language identification accuracy is the percentage of utterances in which the system's top choice language candidate is correct. The rank order statistic is the average position of the correct language within the ordered list of language candidates. The rank order statistic conveys more information about the system's performance than the language identification accuracy and as such is the more prevalent of the two statistics used in this thesis.

3.2 General System Architecture

For this thesis, the system is structured around the segment-based probabilistic framework described in Chapter 2. A system which utilizes this segment-based framework can be realized as a series of three components. These components are a preprocessor, a phonetic recognizer, and a language identifier. The preprocessor receives the raw acoustic waveform as its input and transforms this input into the frame-based feature vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{f}}$. The phonetic recognizer receives the vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{f}}$ as its input and finds the best phonetic hypothesis and segmentation, $\hat{C}$ and $\hat{S}$. The language identifier then uses $\vec{\mathbf{a}}$, $\vec{\mathbf{f}}$, $\hat{C}$ and $\hat{S}$ to find the most likely language. This architecture is displayed in Figure 3.1.

[4] Although the system's design does not consider any issues other than reliable language identification, evaluations of the system's training set requirements and receiver-operator characteristics are presented in this thesis in Chapter 4.

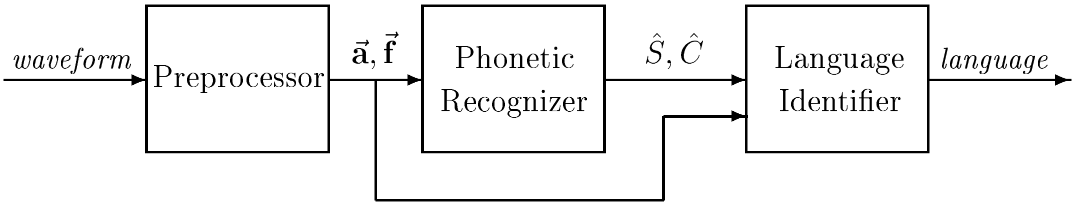

Figure 3.1: System Architecture

3.3 Preprocessing

3.3.1 Spectral Representation

For this thesis, the acoustic vector $\vec{\mathbf{a}}$ is represented with mel-frequency scale cepstral coefficients (MFCC's) [26]. A set of fourteen MFCC's are computed for each utterance with a frame rate of 200 frames per second, a discrete Fourier transform (DFT) size of 256, and a Hamming window of length 25.6 milliseconds. In addition to the MFCC's, fourteen delta MFCC's are also computed. The delta MFCC's are computed with the expression

$$\dot{x}[i] = \frac{1}{2}x[i+1] - \frac{1}{2}x[i-1] \tag{3.1}$$

where $x[i]$ and $\dot{x}[i]$ represent an MFCC value and delta MFCC value for the i^{th} frame. The MFCC signal representation was chosen because it has proven to be an effective representation for speech recognition in various different languages including English [25], Italian [5] and Japanese [12].

3.3.2 Voicing Information

For this thesis, the voicing information contained in the vector $\vec{\mathbf{f}}$ is extracted from the acoustic signal with the *formant* program contained in Entropic's ESPS package. The fundamental frequency tracker contained in the *formant* program is based on an algorithm devised by Secrest and Doddington [33]. The frame rate for $\vec{\mathbf{f}}$ is also 200 frames per second. For each frame, a fundamental frequency (F0) and a probability of voicing parameter are estimated. In an attempt to eliminate speaker dependencies a two step transformation is applied to the F0 values. First, the logarithm (base 2) of F0 is taken for all voiced frames (i.e. frames whose voicing probability is greater than .5). Second, in the logarithm domain, the mean F0 value for each utterance is computed and subtracted from each F0 value. Additionally, a delta F0 value is calculated (also in the logarithm domain) for each voiced frame in the same fashion as the delta MFCC values are found from the MFCC's (see (3.1)).

3.4 Phonetic Recognition

3.4.1 Overview

As previously stated, the fact that the OGI corpus is unlabeled prevents the use of a phonetic recognizer which is trained in a fully supervised manner. It is thus necessary to devise phonetic recognition schemes which do not rely upon fully supervised training. Two possible alternatives that are investigated in this thesis are:

- To train a phonetic recognizer in an unsupervised fashion.
- To train a phonetic recognizer using an alternate database which is labeled.

In developing either of these approaches three main issues must be addressed. These issues are:

- How will the segmentation probability $\Pr(S \mid \vec{\mathbf{a}}, \vec{\mathbf{f}})$ be modeled?
- How will the phonetic classification probability $\Pr(C \mid S, \vec{\mathbf{a}}, \vec{\mathbf{f}})$ be modeled?
- What will the set of phonetic units be?

3.4.2 Phonetic Recognition Utilizing Unsupervised Training

Determining the Best Segmentation

In segment-based approaches, a model for segmentation must be defined. One approach to modeling the probability $\Pr(S \mid \vec{\mathbf{a}}, \vec{\mathbf{f}})$ is to model the probability of the existence of the boundaries which define the segmentation. This approach is used in segment-based approaches such as the stochastic explicit-segment modeling approach proposed by Leung et al. [22]. Because of the tremendous number of possible segmentations which can exist, it is desirable to limit the segmentation search space to a small subset of likely segmentations. One means of accomplishing this search space reduction is to use a hierarchical segmentation algorithm such as the one developed by Glass [8, 9]. In Glass's approach, a dendrogram produced from the spectral information of the signal provides a well organized segmentation search space. The dendrogram is produced by a hierarchical clustering algorithm which clusters segments that are adjacent in time using an acoustic similarity measure.

For the unsupervised approach, the search for the best segmentation $\hat{S}$ will be considered independent of the search for the best linguistic sequence $\hat{C}$. Thus the search for the best segmentation $\hat{S}$ can be represented with the expression

$$\max_{\mathbf{S}} \Pr(S \mid \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{3.2}$$

Since the corpus is not labeled the actual segments within the training data are not known. This makes it impossible to develop an automatic method for finding the single best segmentation which is trained in a supervised manner. Thus, a different means for finding the best segmentation must be devised. One possible way to select a single segmentation is to set a threshold on the acoustic similarity measure used in the dendrogram. This threshold would allow two adjacent segments to be clustered together into one segment only if their acoustic similarity exceeds the threshold. The final segmentation $\hat{S}$ is the segmentation that exists when none of the adjacent clusters have an acoustic similarity exceeding the threshold. For this thesis, the threshold was selected by examining the segmentation output of training utterances in the OGI database, and compromising on a value which limits segment boundary deletions at the expense of increased segment boundary insertions.

Determining the Set of Phonetic Classes

For an approach which utilizes unsupervised training, an automatic method for determining the set of phonetic elements must be used. One simple means for achieving this is to use an unsupervised clustering algorithm. For this thesis, the k-means clustering algorithm is used [4]. The algorithm clusters segments extracted from the training data based on similarity of their acoustic feature vectors. The segment-based acoustic feature vector in this case consists of 14 MFCC values averaged over the length of the segment. The entire set of segment-based feature vectors in the training set are rotated using principal component analysis. The vectors are then scaled by the inverse covariance matrix of the entire set of vectors. The rotation and scaling transforms the original vectors into a set of vectors which contain statistically independent components where each component has a variance of one. The k-means algorithm then utilizes a Euclidean distance metric in its iterative clustering procedure.

When using the k-means algorithm for this purpose, the hope is that each cluster provided by the algorithm will approximately correspond to a specific broad phonetic class (i.e. vowel, fricative, nasal, etc.). Figure 3.2 shows the average Mel frequency spectral coefficient (MFSC) values for each of the clusters found from the k-means algorithm for an experiment where the number of clusters was set to four. As can be seen in the figure, the clusters vary predominately in their energy and do not have extremely distinctive spectral shapes. Unfortunately, this empirical evidence suggests that the clustering algorithm does not provide clusters for the OGI corpus which adequately correspond to broad phonetic classes.[5] Nevertheless, the clustering algorithm was used to create a series of codebooks where the number of entries in the codebooks was varied from 2 to 58.

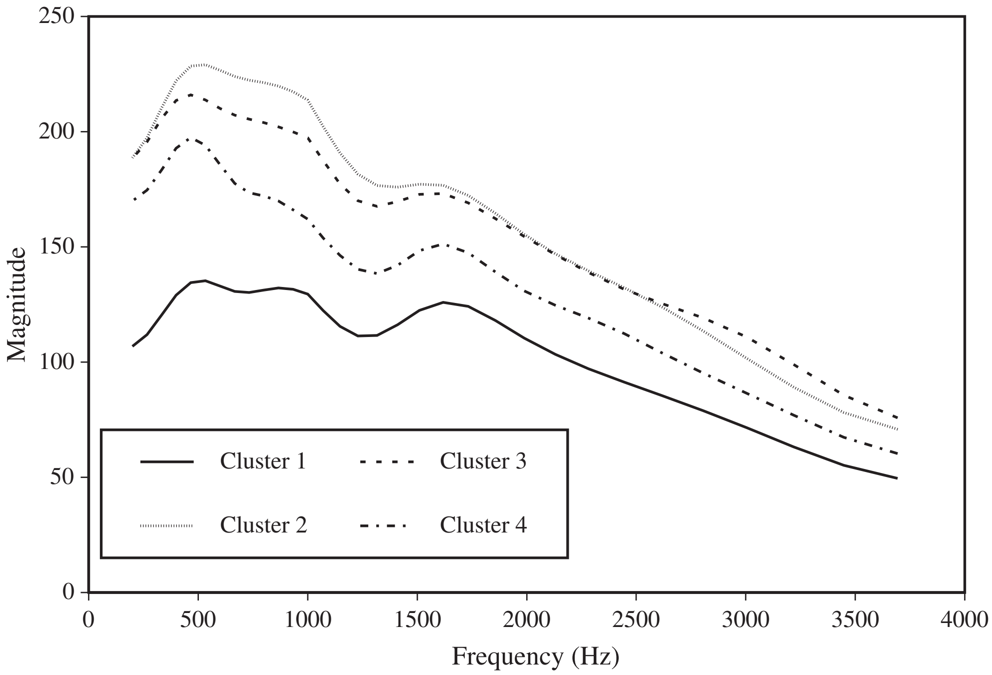

Figure 3.2: Average MFSC values for 4 clusters found with the k-means algorithm

Determining the Best Phonetic String

When the k-means algorithm is used to find a codebook of phonetic units, phonetic classification is performed using vector quantization (VQ) [24]. The use of VQ provides a simple method for modeling the probability $\Pr(C \mid S, \vec{\mathbf{a}}, \vec{\mathbf{f}})$ which is used in determining the string $\hat{C}$. For each segment, the VQ algorithm simply chooses the one codebook entry which most closely matches the acoustic feature vector for that segment. In essence, this is equivalent to assigning a probability of one to the most similar codebook entry and a probability of zero to all other codebook entries.

[5] It should be noted that it may be possible to generate codebooks whose entries more closely resemble broad phonetic classes by using a more sophisticated clustering algorithm than the one presented here. An approach which ignores the energy of each segment may also be preferable.

3.4.3 Phonetic Recognition Utilizing an Alternate Database

The NTIMIT Database

A second possible alternative to completely supervised training is to train a phonetic recognizer on data from an alternate database. One alternative database that could be used is the NTIMIT corpus [14]. NTIMIT contains the utterances from the TIMIT corpus passed through a telephone network [17, 43, 44]. The use of the NTIMIT corpus for training could cause problems for the phonetic recognizer for two reasons. First, the microphones used in collecting the TIMIT data and the phone line channel that the data was passed through may be quite different from the telephone microphones and channels used by the subjects in the OGI corpus. While it is possible that the acoustic differences between the NTIMIT data and the OGI data could be significant, without phonetic transcriptions for the OGI data, it is not possible to quantitatively measure how these differences affect the reliability of the phonetic recognizer when it is used on the OGI data. The second problem associated with training the recognizer using the NTIMIT data is that the NTIMIT corpus only contains utterances collected in English. Because the phones used in English do not comprise the full set of phones used across all the languages in the OGI corpus, highly accurate phonetic recognition can not be achieved. However, it is hoped that, despite the differences between the phone sets of each language, the phonetic labels used in NTIMIT can be collapsed into broad phonetic classes that generalize well across all languages. If this is the case then an accurate multi-language broad phonetic class recognizer may be trained using data from only the English language.

The SUMMIT Phonetic Recognizer

SUMMIT is a segment-based speech recognition system which was developed by the Spoken Language Systems Group at MIT. SUMMIT utilizes Glass's hierarchical segmentation algorithm to provide the segmentation search space. An ordered list of potential phoneme candidates and their respective likelihoods are produced for each potential segment. The phoneme likelihoods are obtained from mixture Gaussian density functions for each phoneme which model segment-based feature vectors. A search algorithm is applied to the segmentation and phoneme search space to find the most likely strings of phonemes. A more detailed description of the SUMMIT system is provided in [37], [40] and [41].

For this thesis, SUMMIT will be used as the phonetic recognition component of the ALI system. To accomplish this, SUMMIT was trained in a fully supervised fashion using the NTIMIT corpus. On NTIMIT, SUMMIT achieved a phonetic recognition accuracy of 60.5 %. Using SUMMIT, the most likely segmentation and string of English phonemes can be found for each utterance in the OGI corpus.

Choosing the Set of Phonetic Units

If the SUMMIT system trained on NTIMIT is used, a means for selecting the set of phonetic classes that will be used must be determined. For this thesis, the number of phonetic classes will be varied to examine the effects of using sets of broad phonetic classes versus sets of more detailed phonetic classes.[6] Strings of broad phonetic classes will be less likely to contain errors than strings using more detailed phonetic elements. However, strings containing more detailed phonetic elements could provide more information if the error rate of the phonetic recognizer is not too overwhelming and there is an adequate amount of training data.

Since the number of phonetic classes used will be varied, a means of determining useful phonetic classes as the number of classes is altered must be devised. One potential means of accomplishing this is to create a phonetic hierarchical structure in which the phonemes are clustered according to a similarity measure. If a phonetic language model is to be used in the language identifier, then a useful similarity measure might compare the contexts in which specific phonemes or phonetic classes appear. By way of example, if two phonemes always appear within similar contexts across all languages, then little detail would be lost by combining the phonemes into one larger class.[7] Figure 3.3 shows a hierarchical phonetic clustering which was obtained by clustering phonemes based upon the contexts in which they appeared in SUMMIT's automatic transcriptions of the training data. Table 3.1 shows the set of phonetic classes that can be extracted from the hierarchical clustering when the number of classes is set to ten.

To obtain the hierarchical phonetic structure in Figure 3.3, clustering was performed in a bottom-up manner. The similarity measure used for the clustering was the divergence between the probability distributions of the different phones. In this case, the distribution for each phone measured the probability of all of the phone's possible left and right contexts. To describe the divergence measure mathematically let $\mathbf{P}$ represents the probability distribution for the expression $\Pr(c_l, c_r \mid c)$. Thus, the distribution $\mathbf{P}$ contains a probability for all possible left and right phonetic contexts, c_l and c_r, for the phone c. Similarly, let $\hat{\mathbf{P}}$ represent the probability distribution $\Pr(c_l, c_r \mid \hat{c})$. Using this notation, the divergence measurement between the distributions $\mathbf{P}$ and $\hat{\mathbf{P}}$ can be expressed as

$$D(\mathbf{P} \| \hat{\mathbf{P}}) = \sum_{c_l, c_r} (\Pr(c_l, c_r \mid c) - \Pr(c_l, c_r \mid \hat{c})) \log \frac{\Pr(c_l, c_r \mid c)}{\Pr(c_l, c_r \mid \hat{c})}. \tag{3.3}$$

[6] The number of classes that can be used has an upper limit of 59. This is the number of distinct phonetic labels that are used by SUMMIT.

[7] Other similarity measures might also prove useful. One possible alternative measure is the acoustic similarity between phones.

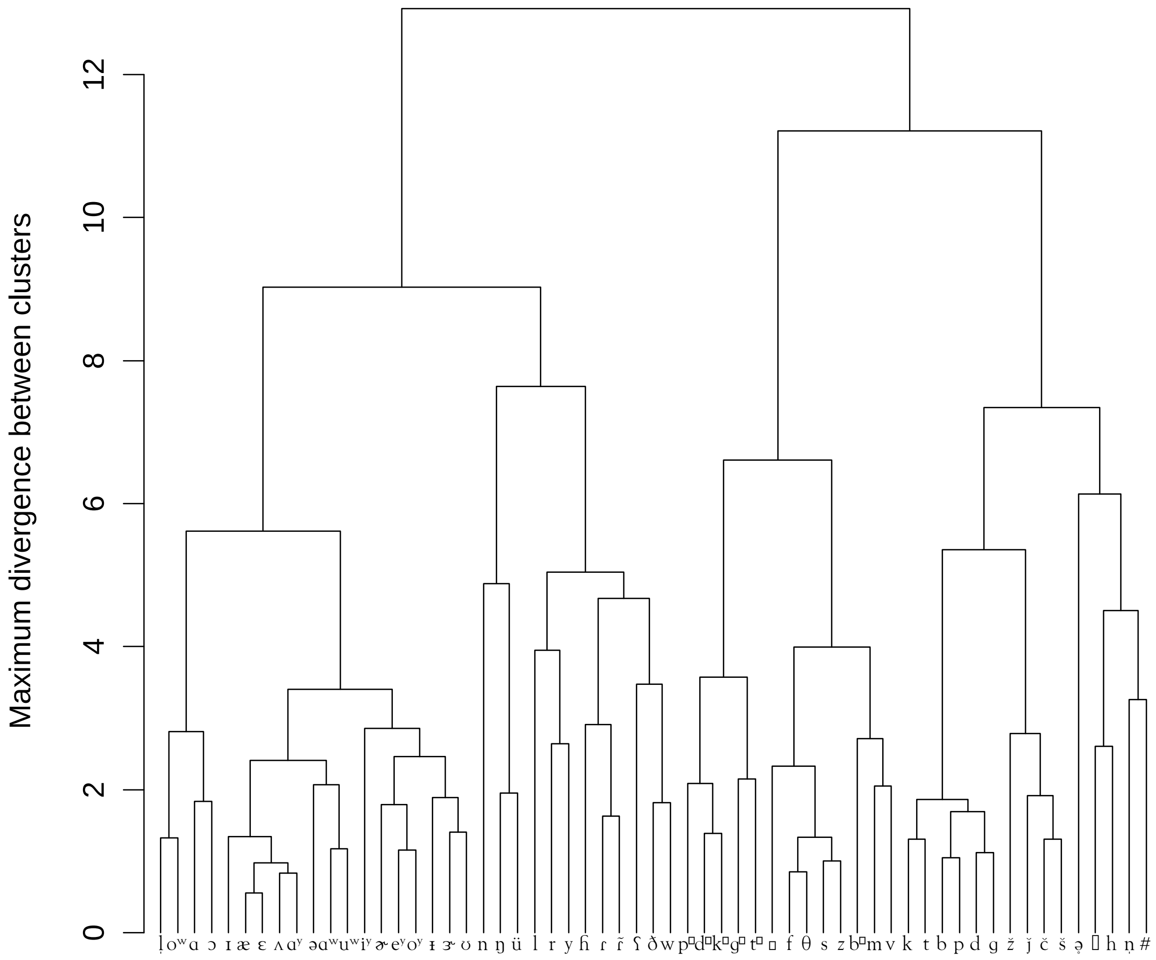

Figure 3.3: Hierarchical clustering of NTIMIT phones into broad phonetic classes

Class	English phonemes in class
1	l̩ oʷ ɑ ɔ
2	iʸ ɪ ɨ eʸ æ ɛ ʌ ɔʸ ɑʸ ɑʷ uʷ ʊ ə ɝ ɚ
3	n ŋ ŋ̍ ü
4	r l y w ɦ ð ɾ ɾ̃ ʔ
5	p□ d□ t□ g□ k□
6	□ f θ s z b□ m m̩ v
7	t d p b k g
8	š č ǰ ž
9	ə̥
10	□ h n̩ #

Table 3.1: Set of ten automatically selected phonetic classes

Class	English phonemes in class
1	# □ □ b□ p□ d□ t□ g□ k□
2	t d p b k g h ɦ m m̩ n n̩ ŋ ŋ̍ ɾ̃ s z θ ð f v ɾ ʔ
3	š č ǰ ž
4	y iʸ ɪ ɨ eʸ ü
5	w uʷ ʊ oʷ ɔ ɔʸ ə l̩ l ə̥
6	ɑ ʌ ɑʸ ɑʷ æ ɛ
7	r ɝ ɚ

Table 3.2: Set of seven manually selected phonetic classes

As can be seen in Figure 3.3 and Table 3.1, the automatic clustering algorithm roughly clusters the phonemes into generic broad phonetic clusters. However, due to sparse data for a few of the phones, such as /ə̥/, /n̩/, and /□/, some of the clusters produced by the hierarchical clustering algorithm are contrary to intuition. Therefore, a number of sets of manually selected broad phonetic classes were also created. These sets of phonetic classes were chosen so that the elements of each class were similar first in their manner of articulation (i.e. vowel, consonant, closure, etc.) and second in their place of articulation (i.e. back, front, etc.). The two sets which proved the most effective in language identification experiments are shown in Tables 3.2 and 3.3.

Class	English phonemes in class
1	# □
2	□
3	b□ p□ d□ t□ g□ k□
4	t d p b k g
5	š ž
6	č ǰ
7	s z
8	θ ð f v
9	ʔ
10	ɾ ɾ̃
11	n n̩
12	ŋ ŋ̩
13	m m̩
14	h ɦ
15	y iʸ eʸ
16	æ
17	ɑ ɔ
18	w uʷ oʷ
19	ɪ ɨ ü ɛ ʌ ʊ ə ə̥
20	ɔʸ ɑʸ
21	ɑʷ
22	l l̩
23	r ɝ ɚ

Table 3.3: Set of 23 manually selected phonetic classes

3.5 Language Identification

3.5.1 Issues

Using the framework discussed in Chapter 2, the language identification component of the system models the expression

$$\arg\max_i \Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i) \Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) \Pr(\hat{C} \mid L_i) \Pr(L_i). \tag{3.4}$$

Thus, the modeling issues involved in the language identification component of the system can be summarized with the following questions:

- How will the *a priori* language probability, $\Pr(L_i)$, be modeled?
- What language model will be used to represent $\Pr(\hat{C} \mid L_i)$?
- What prosodic model will be used to represent $\Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i)$?
- What acoustic model will be used to represent $\Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i)$?

3.5.2 *A Priori* Language Probability

The *a priori* language probability, $\Pr(L_i)$, is perhaps the simplest element in the system to model. The only concern is determining how the language probabilities should be estimated. One potential solution is to say that all of the candidate languages are equally likely to be spoken. In effect this is the assumption that is made for the maximum likelihood approach to ALI. A different approach would be to attempt to estimate the probability of encountering each candidate language in the environment where the ALI system is to be used. Because the OGI corpus contains nearly equal amounts of data from each language, we assume that each candidate language is equally likely to be spoken. With this assumption the term $\Pr(L_i)$ can simply be ignored in the language identification process.

3.5.3 Language Model

Overview

The language model is used to represent the expression $\Pr(\hat{C} \mid L_i)$. The language model is potentially the most important element of the system. As House and Neuburg showed, simple language models applied to error free sequences of broad phonetic classes can reliably identify the language of an utterance. In this thesis, an n-gram language model is investigated. More specifically, the unigram, bigram, and trigram model are examined independently, as well as in combination.

Basic n-gram Modeling

Some simple assumptions are made in the derivation of the n-gram model. For a unigram model each phonetic element is assumed to be statistically independent of all other phonetic elements. This can be expressed mathematically as

$$\Pr(\hat{C} \mid L_i) = \Pr(c_1, c_2, \ldots, c_p \mid L_i) = \prod_{k=1}^{p} \Pr(c_k \mid L_i). \tag{3.5}$$

A bigram model assumes each phonetic element is statistically dependent on only the phonetic element immediately preceding it. This is expressed mathematically as

$$\Pr(\hat{C} \mid L_i) = \Pr(c_1 \mid L_i) \prod_{k=2}^{p} \Pr(c_k \mid c_{k-1}, L_i). \tag{3.6}$$

Similarly, a trigram model assumes each linguistic element is statistically dependent on the two preceding phonetic elements. This is expressed as

$$\Pr(\hat{C} \mid L_i) = \Pr(c_1 \mid L_i) \Pr(c_2 \mid c_1, L_i) \prod_{k=3}^{p} \Pr(c_k \mid c_{k-1}, c_{k-2}, L_i). \tag{3.7}$$

To utilize these models for language identification, the probabilities for each language dependent n-gram model are estimated from histogram counts for each phonetic element. The histograms are generated from the phonetic labels attached to the training utterances by the phonetic recognizer. To avoid the possibility of having probabilities of zero within the n-gram models, each histogram is initialized with an arbitrarily chosen minimum count floor of $\frac{1}{p}$ where p is the number of phonetic classes.

In evaluating the performance of the basic n-gram model there are four considerations that must be taken into account. These considerations are summarized in the following questions:

- How accurately does $\hat{C}$ represent the underlying string of phonetic elements?
- How many phonetic classes are used to represent the elements in $\hat{C}$?
- What is the value of n for the n-gram model?
- How much training data is being used?

The performance of the n-gram model is extremely dependent on the four issues stated above.

The language model component of the system attempts to capture the phonotactic constraints of each of the languages using the n-gram statistics of $\hat{C}$. In order to do this properly, it is important that $\hat{C}$ represent the actual string of phonetic events as accurately as possible. House and Neuburg showed that the phonotactic constraints of languages are so strong that accurate language identification can be performed using simple n-gram models even when the string of phonetic events is modeled with elements as general as broad phonetic classes. However, the language identification capabilities of an n-gram will be degraded when the actual string of phonetic events is corrupted with errors.

The introduction of errors into the phonetic string has one major consequence with respect to n-gram modeling. As the phonetic recognition error rate within $\hat{C}$ is increased, the probability distributions within the n-gram models are shifted away from their actual distributions towards more uniform distributions. This shifting of the n-gram probability distributions towards more uniform distributions decreases the language discrimination abilities of the n-gram model. It should also be noted that this effect becomes greater as the size of the n-gram model is increased. This can be attributed to the fact that more past information must be used as the value of n is increased. In other words, when n is greater than one, the n-gram model is subjected not only to errors in the current phonetic element but also to errors in the previous phonetic elements that are used for the context dependency of the model.

To examine the effect the inventory of phonetic elements used in the phonetic string has upon the n-gram model performance, the n-gram model was tested using the two different methods for determining $\hat{C}$ that were described earlier (i.e., the SUMMIT phonetic recognizer and the vector quantizer). When the SUMMIT phonetic recognizer was used, the string of phonetic classes was determined by collapsing the detailed labels produced by SUMMIT into the phonetic classes produced by the hierarchical clustering shown in Figure 3.3. Figure 3.4 shows the language identification accuracy of the unigram model using the phonetic string output of both the SUMMIT phonetic recognizer and the vector quantizer as the number of phonetic classes is varied from 2 to 59. The accuracy is shown for both the training and test sets. Figure 3.5 shows the rank order statistic for the same set of experiments.

As can be seen in Figure 3.4 and Figure 3.5 the unigram model using the SUMMIT supplied phonetic string outperforms the unigram model using the vector quantizer's phonetic string as the number of phonetic classes is increased beyond 10. This is expected since the SUMMIT recognizer provides a more accurate phonetic representation of the utterance than the vector quantizer. However, when the number of classes is less than ten, the unigram model performs better with the vector quantizer than with SUMMIT. Figures 3.6 and 3.7 show the same experiments using a bigram model instead of a unigram model. Similar to the unigram model, the bigram model performs better using the SUMMIT recognizer than it does using the vector quantizer when the number of phonetic classes is selected to be greater than 7.

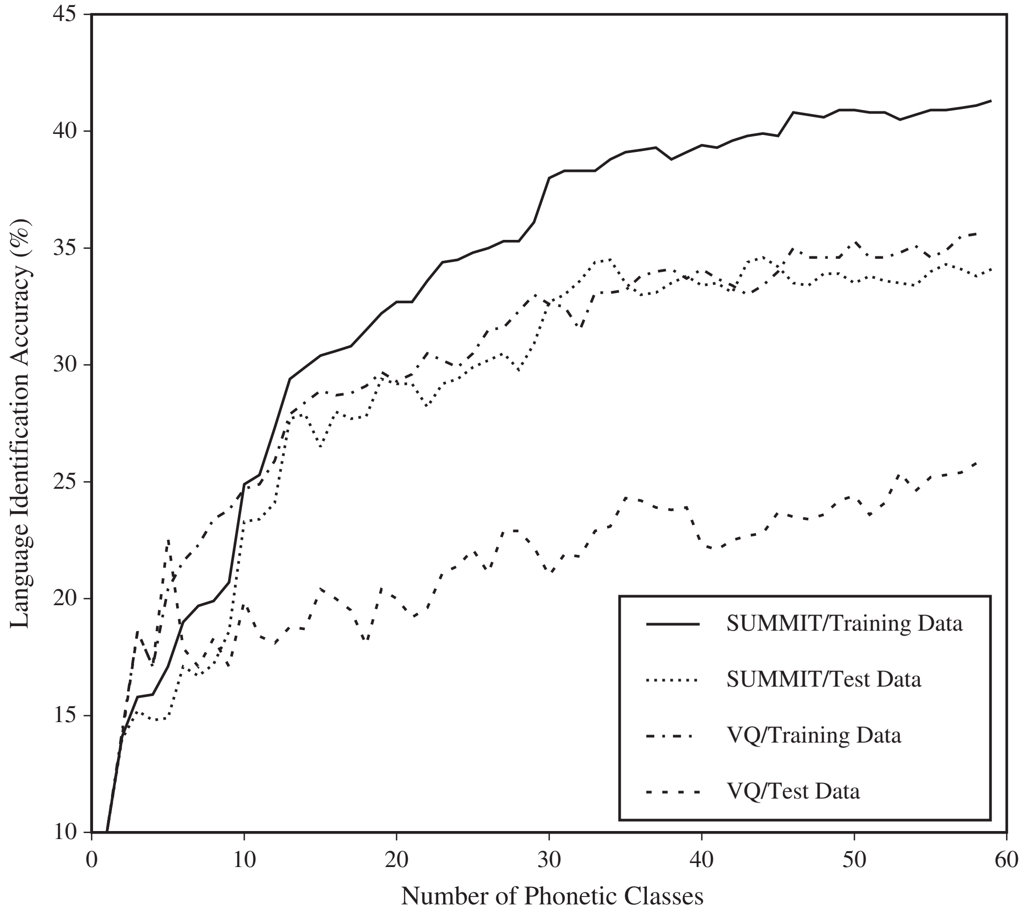

Figure 3.4: Accuracy of unigram model using two different phonetic recognizers as the number of phonetic classes is varied

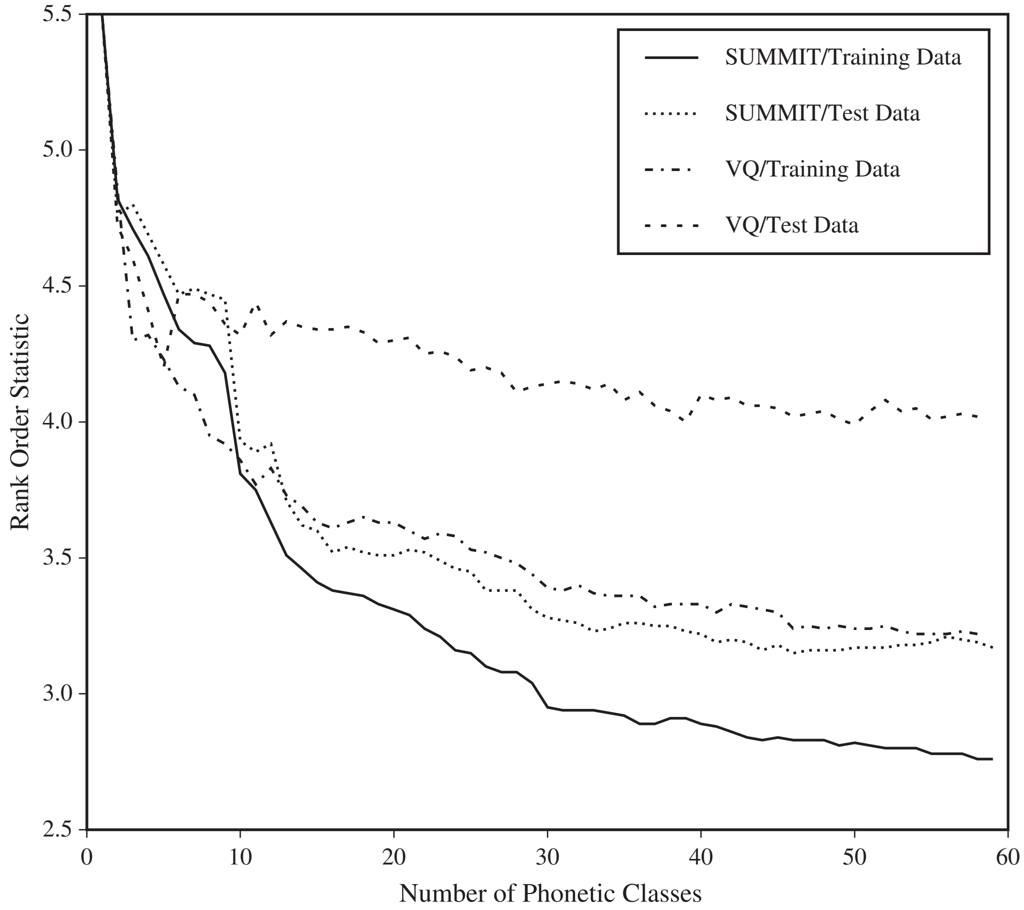

Figure 3.5: Rank order statistic of unigram model using two different phonetic recognizers as the number of phonetic classes is varied

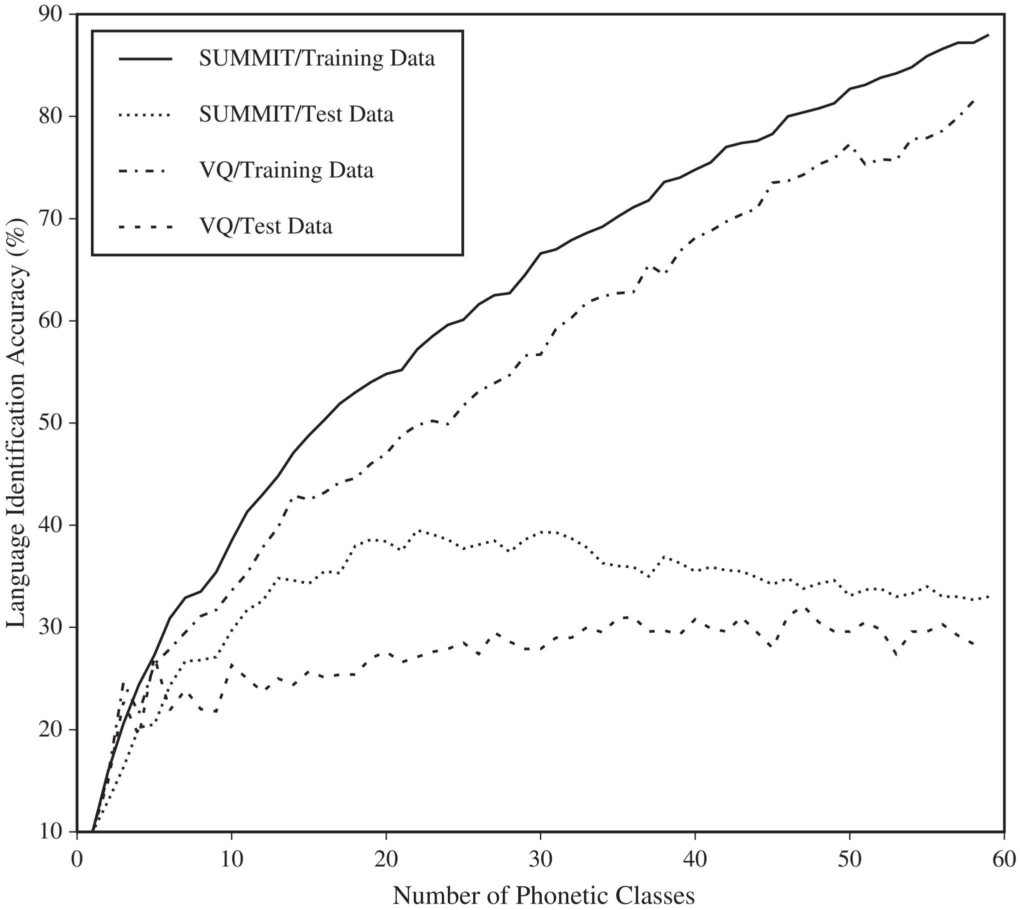

Figure 3.6: Accuracy of bigram model using two different phonetic recognizers as the number of phonetic classes is varied

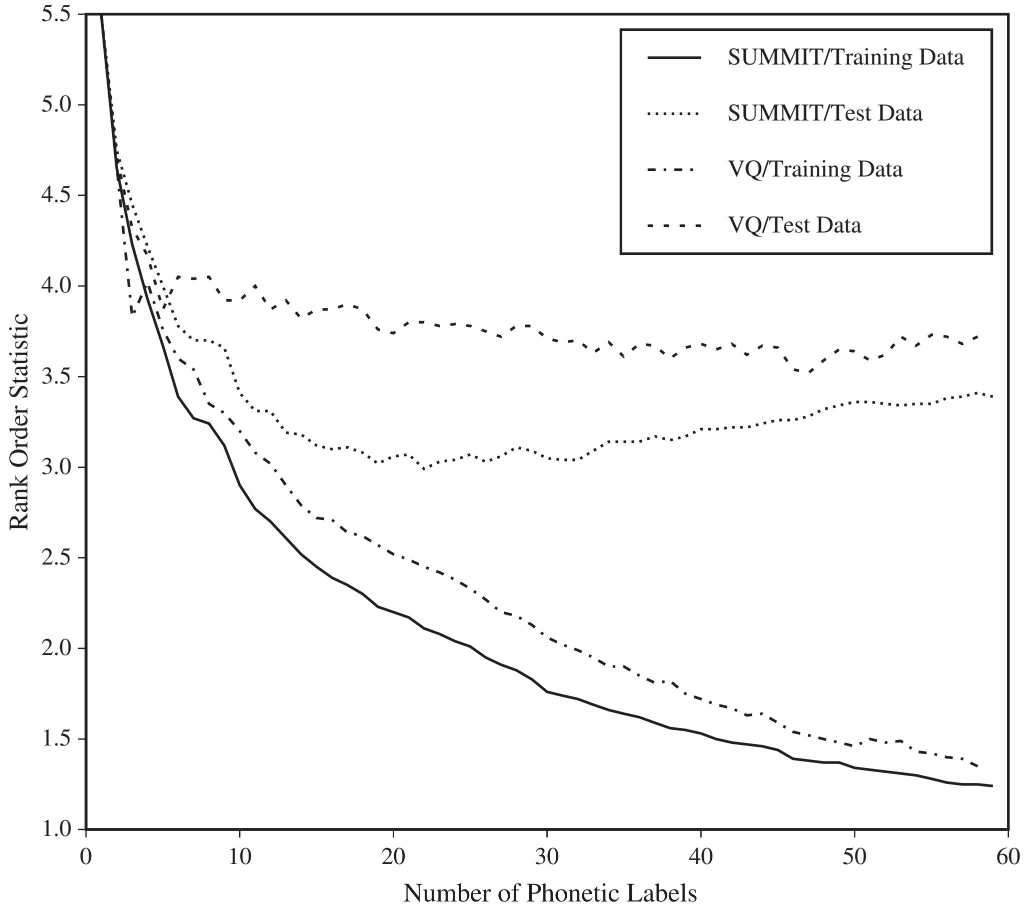

Figure 3.7: Rank order statistic of bigram model using two different phonetic recognizers as the number of phonetic classes is varied

n-gram Model	Determination of Classes	Number of Classes	Language ID Accuracy	Rank Order Statistic
Bigram	Automatic	22	39.5	2.99
	Manual	23	41.5	2.95
Trigram	Automatic	7	27.4	3.70
	Manual	7	34.8	3.27

Table 3.4: Performance of n-gram models using automatically and manually selected broad phonetic classes obtained from the phonetic labels provided by SUMMIT

To further examine the effect the representation of $\hat{C}$ has on the language model performance, several experiments were also conducted using manually selected broad phonetic classes instead of the automatically selected classes produced by the hierarchical clustering. Table 3.4 compares the performance of the best bigram and trigram models using the automatically selected phonetic classes with the performance of the bigram and trigram models using the manually selected classes shown in Tables 3.2 and 3.3. As can be observed in Table 3.4, the performance of n-gram models was better using the manually selected classes than the automatically selected classes. These results further demonstrate the importance of using meaningful phonetic classes in the representation of the phonetic string.

The amount of available training data is also a primary concern. The size of the n-gram model and the number of phonetic classes should be chosen to provide as much detail as possible. However, increasing the detail of the modeling also increases the amount of data required for proper training. In general, as the number of parameters in the n-gram model is increased, the training requirements of the model are also increased. If n is the size of the n-gram model and p is the number of phonetic classes in the phonetic string, then the n-gram model for each language contains p^n parameters that must be estimated. Thus, as either n or p is increased, the detail in the n-gram model is increased, thereby increasing the amount of data that is needed for proper training.

The tradeoff between the increased detail and the increased training requirements can be observed in Figures 3.8 and 3.9. For small numbers of phonetic classes, the trigram model outperforms the bigram and unigram models. This is expected since the trigram provides longer distance constraints in modeling the phonetic string than the bigram and unigram models. However, as the number of classes is increased the performance of the trigram drops off severely due to the lack of sufficient amounts of data to properly train the large number of parameters in the trigram model. A similar effect is seen between the performance of the bigram and unigram models.

The bigram model easily outperforms the unigram when the number of classes is less than 40. However, as the number of classes is increased, the bigram experiences the same drop in performance as the trigram model due to insufficient training data. As the number of classes is increased above 50, the unigram model begins outperforming both the bigram and trigram models, as it has considerably fewer parameters that need to be trained.

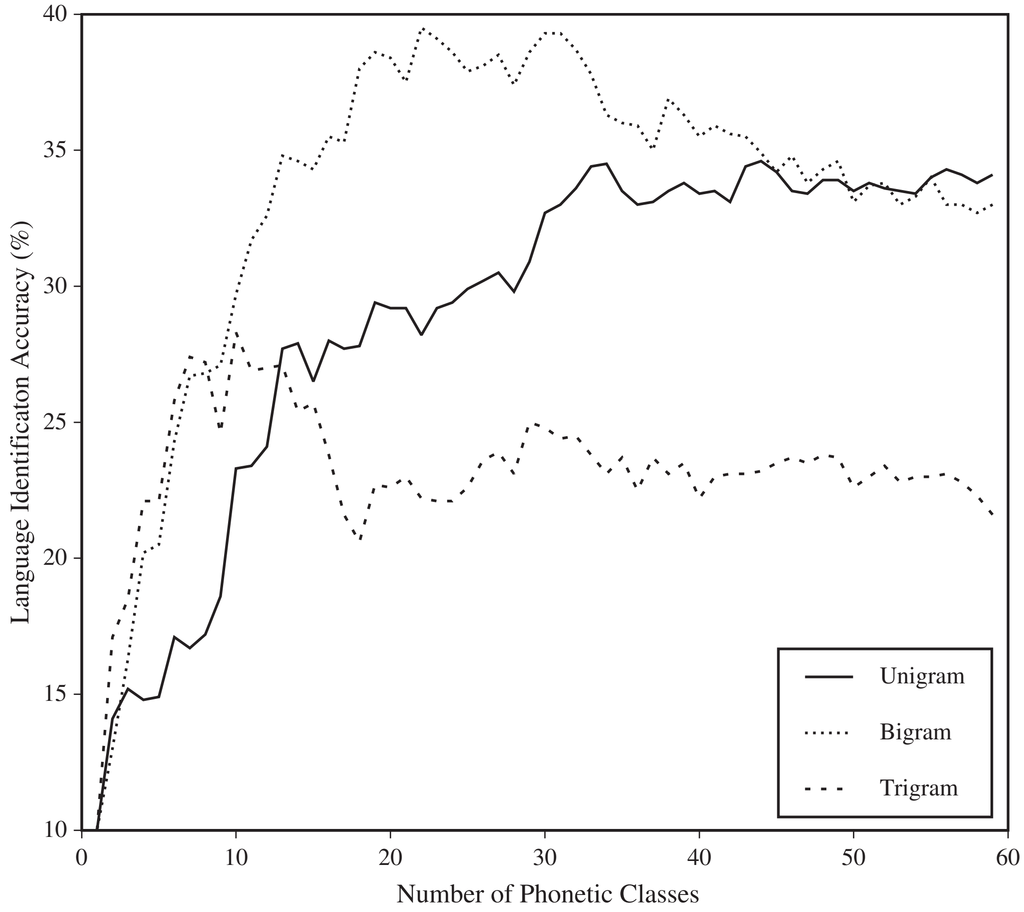

Figure 3.8: Language identification accuracy of n-gram models using the SUMMIT phonetic recognizer with automatically selected classes as the number of phonetic classes is varied

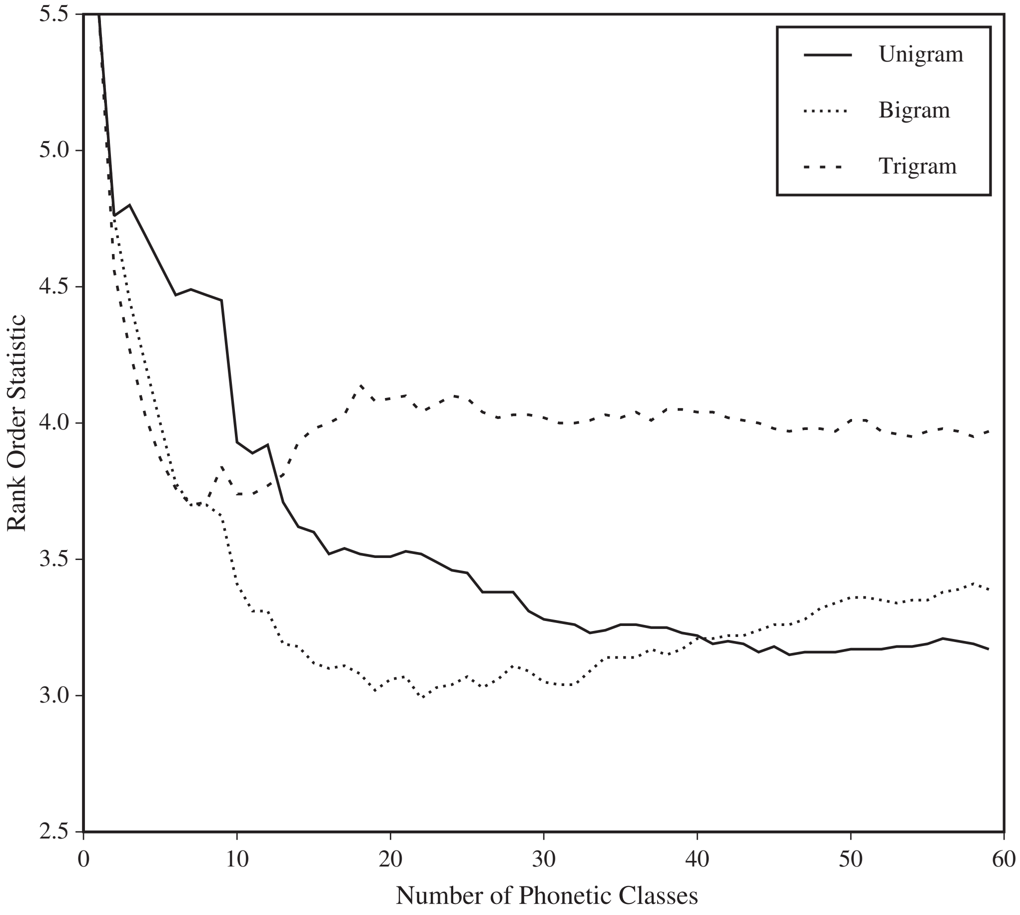

Figure 3.9: Rank order statistic of n-gram models using the SUMMIT phonetic recognizer with automatically selected classes as the number of phonetic classes is varied

To examine the training requirements of each of the n-grams models in more detail, each n-gram model was tested using varying training set sizes. Figures 3.10, 3.11 and 3.12 show how the performance of the unigram, bigram and trigram models varies as the training set size is increased from 10 speakers per language to 50 speakers per language. An examination of Figure 3.10 reveals that very little improvement in performance is likely to be gained in the unigram model by simply increasing the number of training speakers beyond 50. However, significant improvements in the bigram and trigram models' performance may be possible with only a moderate increase in the number of training speakers.

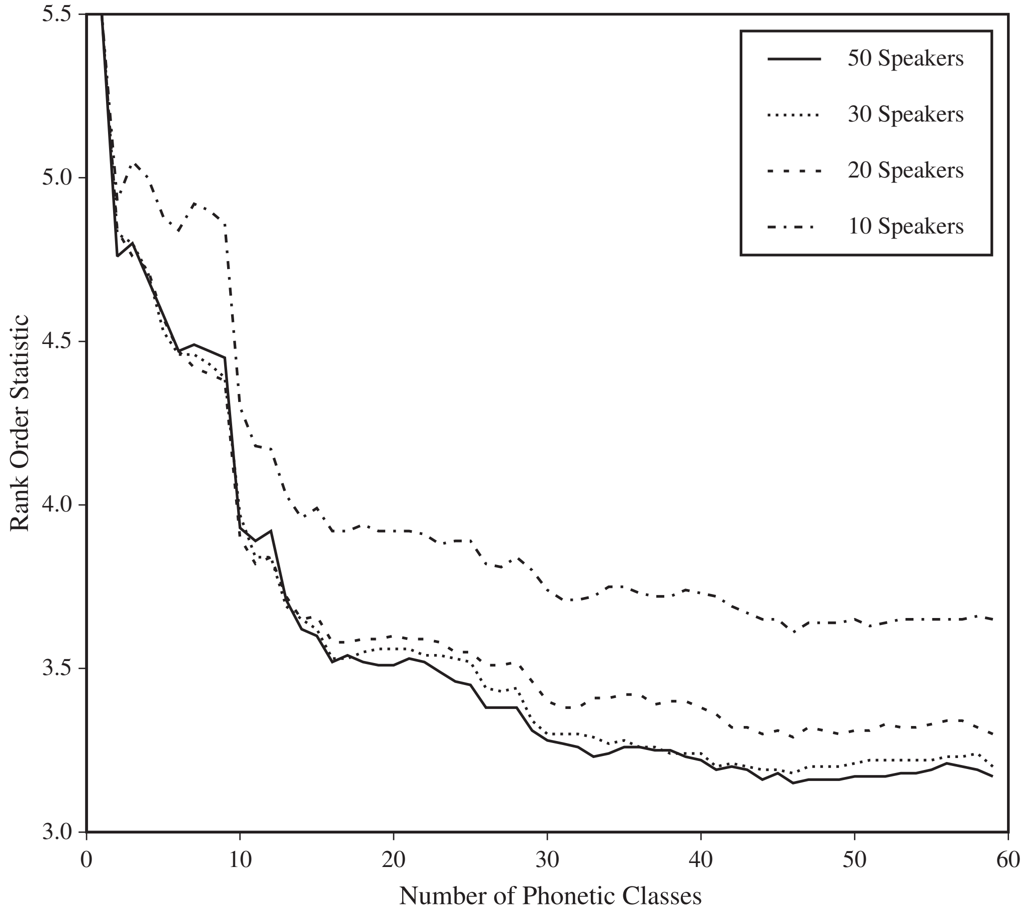

Figure 3.10: Performance of unigram model using the SUMMIT recognizer with automatically selected classes as the number of training speakers per language is altered

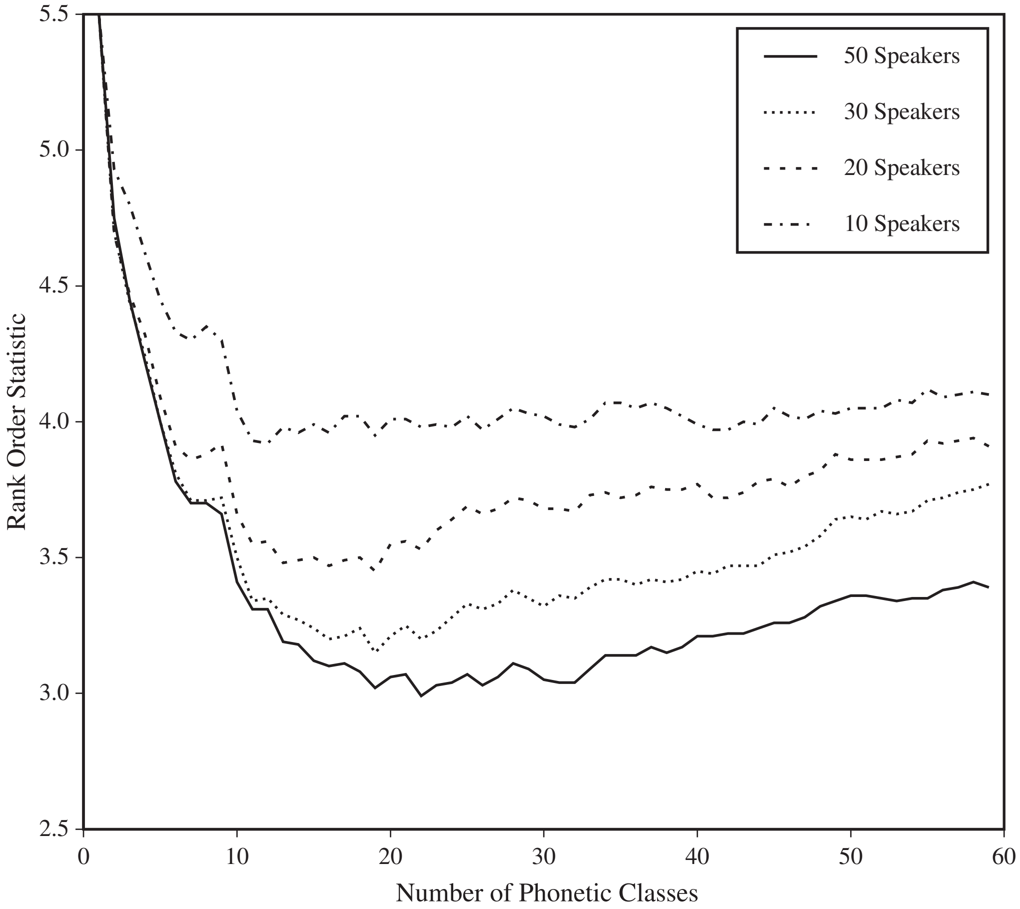

Figure 3.11: Performance of bigram model using the SUMMIT recognizer with automatically selected classes as the number of training speakers per language is altered

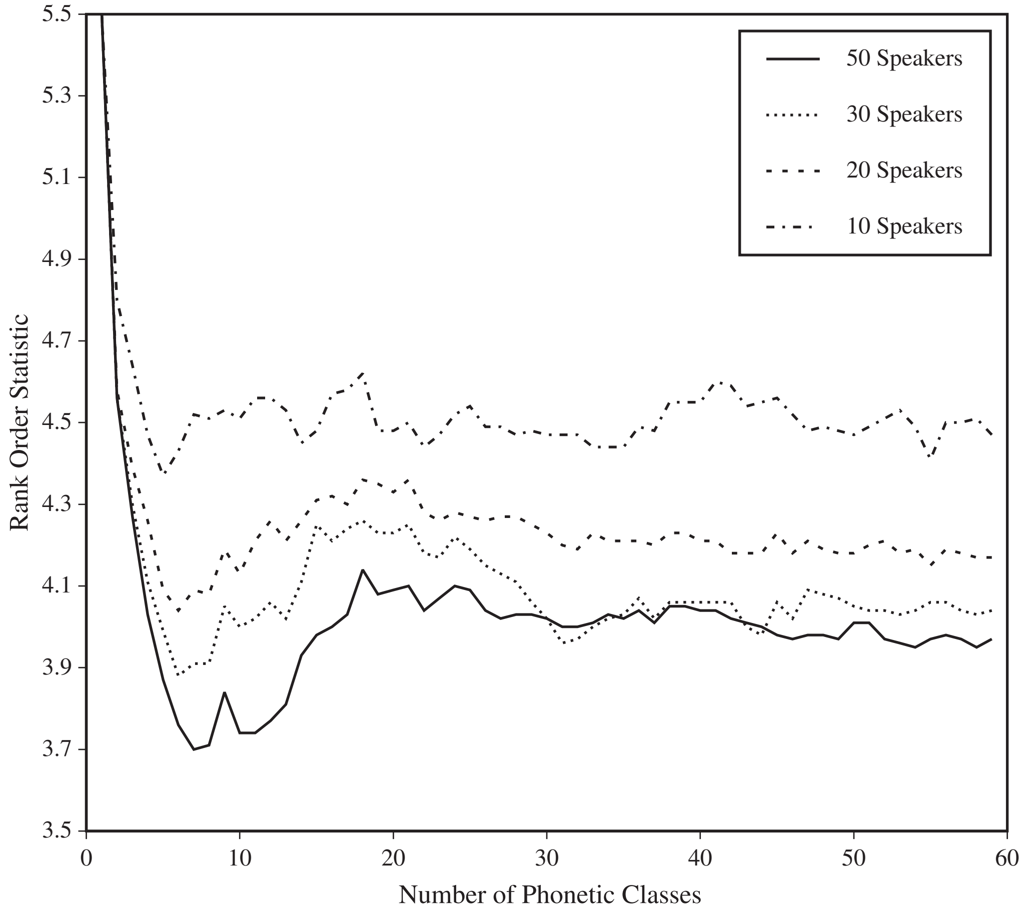

Figure 3.12: Performance of trigram model using the SUMMIT recognizer with automatically selected classes as the number of training speakers per language is altered

Interpolated n-gram Modeling

One means of reconciling the tradeoff between the advantage of increased detail and the disadvantage of limited training data as the number of phonetic classes is increased is to utilize an interpolation approach for combining the different n-gram models [15]. The idea behind interpolation is to utilize the strength of larger n-gram models when sufficient training data is available for specific contexts but to rely more heavily on smaller n-gram models when the training data available for a specific context is limited.

The simplest interpolated n-gram model is the interpolated bigram model. The interpolated bigram is described by the expression

$$\hat{\mathrm{P}}(c_i \mid c_{i-1}) = \lambda \Pr(c_i \mid c_{i-1}) + (1-\lambda)\Pr(c_i). \tag{3.8}$$

In (3.8) the interpolated bigram probability $\hat{\mathrm{P}}(c_i \mid c_{i-1})$ is modeled as a linear interpolation of the estimated bigram probability, $\Pr(c_i \mid c_{i-1})$, and the estimated unigram probability $\Pr(c_i)$. The interpolation factor λ is chosen to place more weight on the estimated bigram when there are enough exemplars of c_{i-1} in the training data to properly estimate the bigram probability. When there are limited exemplars of c_{i-1} in the training set, λ shifts the weight onto the estimated unigram probability. The formula for the interpolation factor is

$$\lambda = \frac{k_{c_{i-1}}}{k_{c_{i-1}} + K} \tag{3.9}$$

where $k_{c_{i-1}}$ is the number of exemplars of c_{i-1} in the training set and K is a constant. Ideally, K should be set to a value which optimizes the performance of the interpolated bigram model on the language identification task.

The interpolated bigram can be expanded to larger interpolated n-gram models in a simple recursive fashion. For example, the interpolated trigram is represented with the expression

$$\hat{\mathrm{P}}(c_i \mid c_{i-1}, c_{i-2}) = \lambda_2 \Pr(c_i \mid c_{i-1}, c_{i-2}) + (1-\lambda_2)\hat{\mathrm{P}}(c_i \mid c_{i-1}) \tag{3.10}$$

which expands to

$$\hat{\mathrm{P}}(c_i \mid c_{i-1}, c_{i-2}) = \lambda_2 \Pr(c_i \mid c_{i-1}, c_{i-2}) + (1-\lambda_2)\left(\lambda_1 \Pr(c_i \mid c_{i-1}) - (1-\lambda_1)\Pr(c_i)\right). \tag{3.11}$$

Tests on data jackknifed from the training set revealed that an appropriate K value for the interpolated bigram is 150. Similarly, a value of 1200 was found to be an appropriate K value for the interpolated trigram. Thus λ_1 can be expressed as

$$\lambda_1 = \frac{k_{c_{i-1}}}{k_{c_{i-1}} + 150} \tag{3.12}$$

and λ_2 can be expressed as

$$\lambda_2 = \frac{k_{c_{i-1},c_{i-2}}}{k_{c_{i-1},c_{i-2}} + 1200}. \tag{3.13}$$

Figure 3.13 shows the performance of the interpolated bigram and trigram models in comparison to the standard unigram, bigram and trigram models. As can be observed in the figure, the interpolated bigram and trigram models outperform the standard n-gram models. Additionally, the interpolated models do not experience any drop in performance as the number of phonetic classes is increased although their performance does level off as the number of phonetic classes is increased beyond 30. When 59 phonetic classes were used the interpolated trigram achieved a language identification accuracy of 41.7% and a rank order statistic of 2.78. It should also be noted that the interpolated trigram offers only a slight improvement in performance over the interpolated bigram.

3.5.4 Prosodic Model

Overview

The prosodic model is used to represent the expression $\Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i)$. Ideally, this model can be used to capture the differences among languages that exist in the prosodic structure of utterances. To accomplish this the model should incorporate knowledge about the manner in which word and sentence level stress are incorporated into utterances as well as the usage of tones. Unfortunately, while useful and reliable methodologies are available for modeling acoustic and phonetic information, well-developed techniques for automatically capturing and understanding prosodic information are not yet available. Therefore, for these experiments, the prosodic model is only used to capture simple statistical information about the fundamental frequency and the segment duration information of an utterance.

To help simplify the modeling, the expression for the prosodic model can be expanded as follows:

$$\Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) = \Pr(\vec{\mathbf{f}} \mid \hat{S}, \hat{C}, L_i) \Pr(\hat{S} \mid \hat{C}, L_i). \tag{3.14}$$

With this expansion the prosodic model can be expressed as the product of two separate models, a fundamental frequency model and a segment duration model.

Fundamental Frequency Model

The expression $\Pr(\vec{\mathbf{f}} \mid \hat{S}, \hat{C}, L_i)$ can be used to capture the information available in the F0 contour of an utterance. Although, there may be correlation between the F0

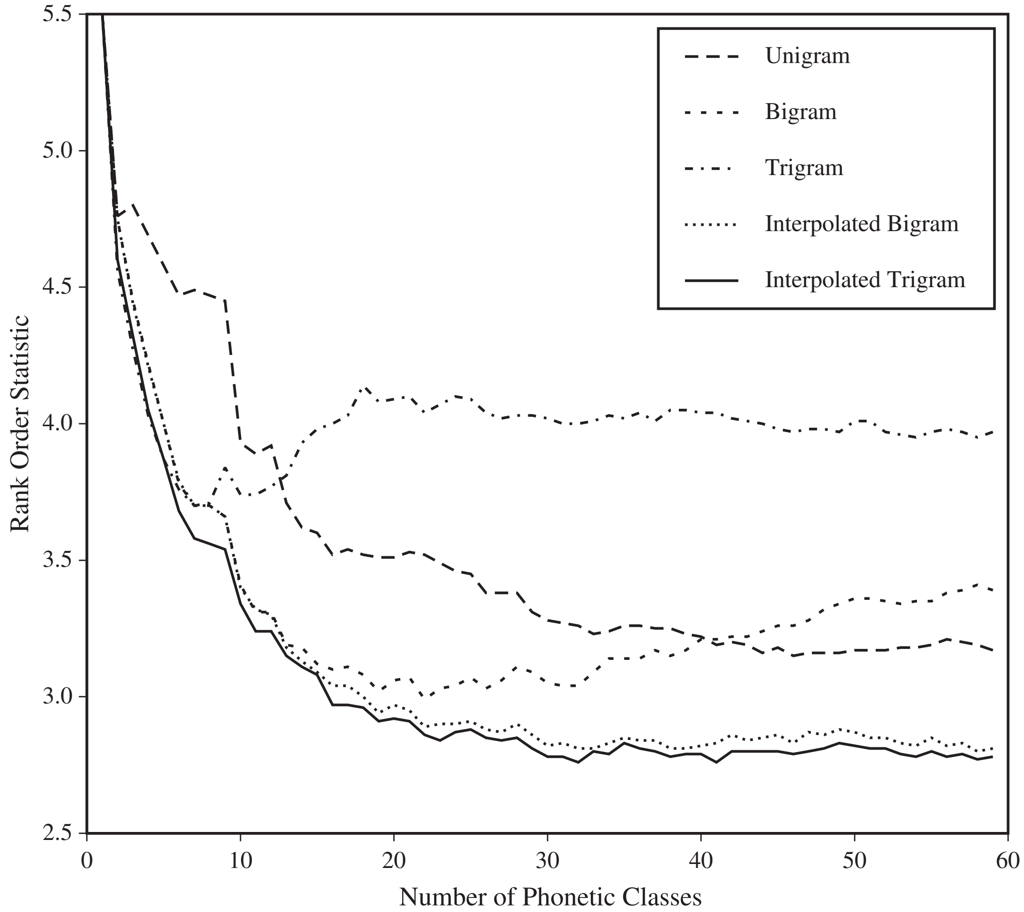

Figure 3.13: Performance of n-gram and interpolated n-gram models as the number of phonetic classes is varied

contour and the durations of the segments in the utterance, this correlation will be ignored for these experiments in order to simplify the modeling of the F0 contour. Thus, $\vec{\mathbf{f}}$ will be considered independent of $\hat{S}$ and $\hat{C}$. With these assumptions the fundamental frequency model can be simplified as follows:

$$\Pr(\vec{\mathbf{f}} \mid \hat{S}, \hat{C}, L_i) = \Pr(\vec{\mathbf{f}} \mid L_i). \tag{3.15}$$

While there may be useful information available in the dynamics of the F0 contour, a method for modeling these dynamics over time for the purpose of language identification is not yet obvious. Some of this dynamic information is presumably captured in the delta F0 values contained in $\vec{\mathbf{f}}$. To simplify the modeling, each frame will be considered to be statistically independent. With this assumption the F0 model can be written as

$$\Pr(\vec{\mathbf{f}} \mid L_i) = \prod_{k=1}^{m} \Pr(\vec{f}_k \mid L_i) \tag{3.16}$$

where m is the number of frames in the utterance and $\vec{f}_k$ is a feature vector representing the F0 and delta F0 values for the k^{th} frame. It should be mentioned that the computation in (3.16) only includes the frames which are voiced.

The expression in (3.16) can be modeled with a mixture of full covariance Gaussian probability density functions. To create the mixture Gaussian model for each language, the set of Gaussians density functions within each mixture are initialized from a set of clusters found with the k-means clustering algorithm. The Gaussians in each mixture are then iteratively reestimated to maximize the average likelihood score of the vectors in the training set.

To find the number of Gaussians within each mixture which is sufficient for modeling the probability density function of the F0 vectors in each language, the performance of the F0 model was examined as the number Gaussians in each mixture was varied from 1 to 24. Figures 3.14 and 3.15 show the performance of the F0 model as the number of Gaussians per mixture is varied. As can be seen, the performance of the model levels off as the number of Gaussians per mixtures is increased to 9 or higher. When 9 Gaussians per mixture are used, the F0 model achieves a language identification accuracy of 23.1% with a rank order statistic of 4.01.

Segment Duration Model

The expression $\Pr(\hat{S} \mid \hat{C}, L_i)$ can be used to capture the segment duration information in a utterance. While there may be very useful information contained in $\hat{S}$ regarding the stress patterns of the syllables, words and sentences in each utterance, this information could require fairly complex modeling and as such will be ignored for

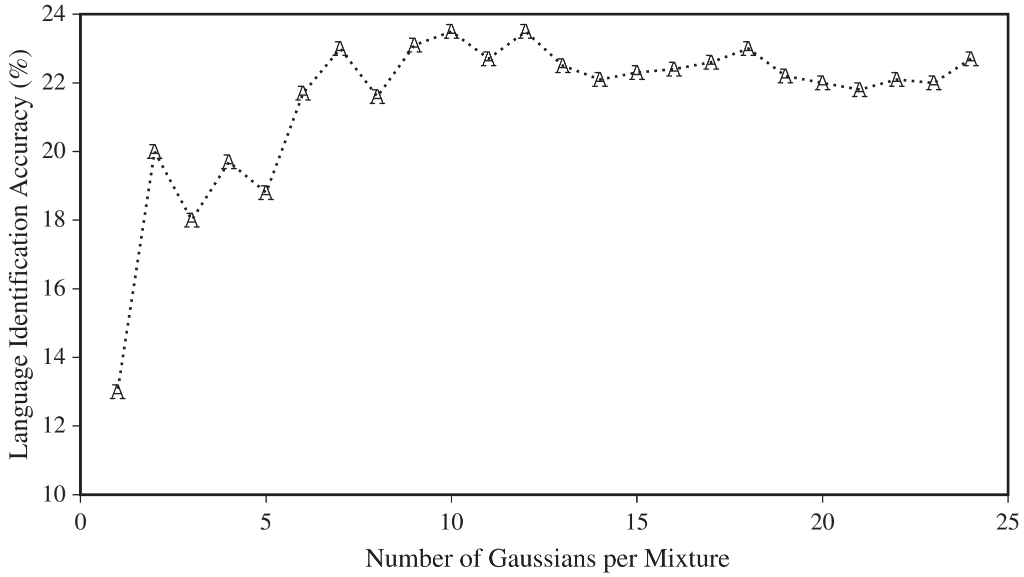

Figure 3.14: Accuracy of F0 model as the number of Gaussians per mixture is varied

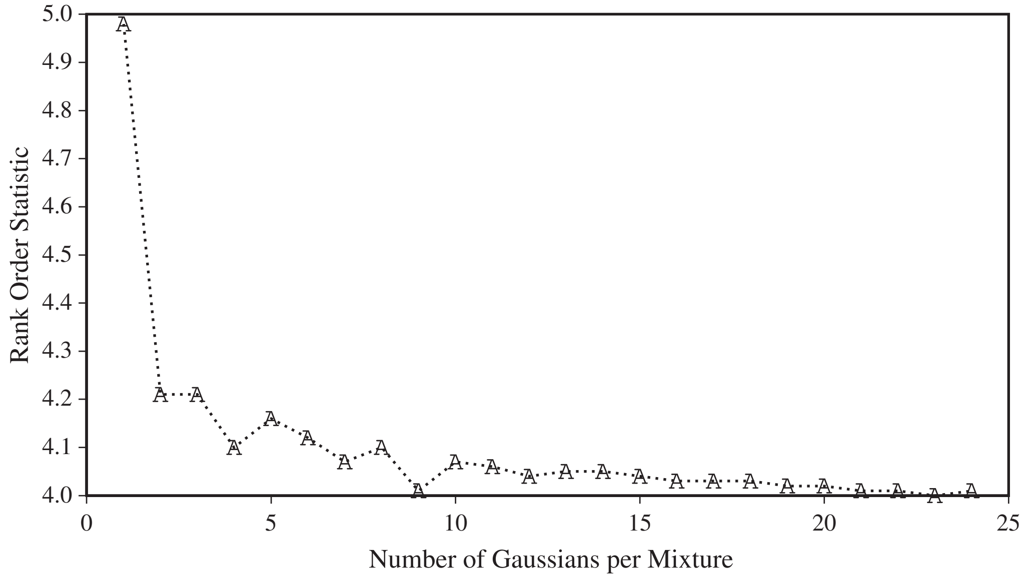

Figure 3.15: Rank order statistic of F0 model as the number of Gaussians per mixture is varied

these experiments in deference to simplicity. As a simplifying assumption each segment will be considered independent of all other segments. With this independence assumption, the segment duration model can be rewritten as

$$\Pr(\hat{S} \mid \hat{C}, L_i) = \prod_{k=1}^{m} \Pr(d_k | c_k, L_i) \tag{3.17}$$

where m is the number of segments in the utterance and d_k is the duration of the k^{th} segment.

For these experiments the duration d_k is expressed as the number of frames contained within the segment. With this consideration, the expression in (3.17) can be modeled directly with non-parametric probability distributions. A probability distribution is created for each phonetic class in each language from histograms which count the number of times specific durations occur in the training data. To help smooth the tail of each histogram (i.e., the histogram bins corresponding to long segment durations), a minimum count floor was applied to each histogram bin and each histogram was smoothed with a low pass filter (i.e., Parzen windowing) before being used to generate the duration probability distributions.

Figures 3.16 and 3.17 show the performance of the segment duration model on the training and test data as the number of phonetic classes is varied from 1 to 59. As can be seen in the figures, the model suffers from insufficient training when larger numbers of classes are used. The peak performance of the segment duration model occurs when 29 phonetic classes are used. With 29 phonetic classes, the model achieves a language identification accuracy of 27.6% and a rank order statistic of 3.65.

3.5.5 Acoustic Model

The expression $\Pr(\vec{\mathbf{a}} \mid \vec{\mathbf{f}}, \hat{S}, \hat{C}, L_i)$ is called the acoustic model. This model is used to capture information about the acoustic realizations of each of the phonetic elements used in each language. To simplify the modeling, the acoustic information $\vec{\mathbf{a}}$ will be assumed independent of the fundamental frequency information $\vec{\mathbf{f}}$. With this assumption the acoustic model can be rewritten as follows:

$$\Pr(\vec{\mathbf{a}} \mid \vec{\mathbf{f}}, \hat{S}, \hat{C}, L_i) = \Pr(\vec{\mathbf{a}} \mid \hat{S}, \hat{C}, L_i). \tag{3.18}$$

To further simplify the expression, each segment will be considered independent of all other segments. With this assumption the acoustic model can be expressed as

$$\Pr(\vec{\mathbf{a}} \mid \hat{S}, \hat{C}, L_i) = \prod_{k=1}^{m} \Pr(\vec{a}_k \mid c_k, L_i) \tag{3.19}$$

where m is the number of segments in the utterance, and $\vec{a}_k$ is a segment-based feature vector describing the acoustics of the k^{th} segment.

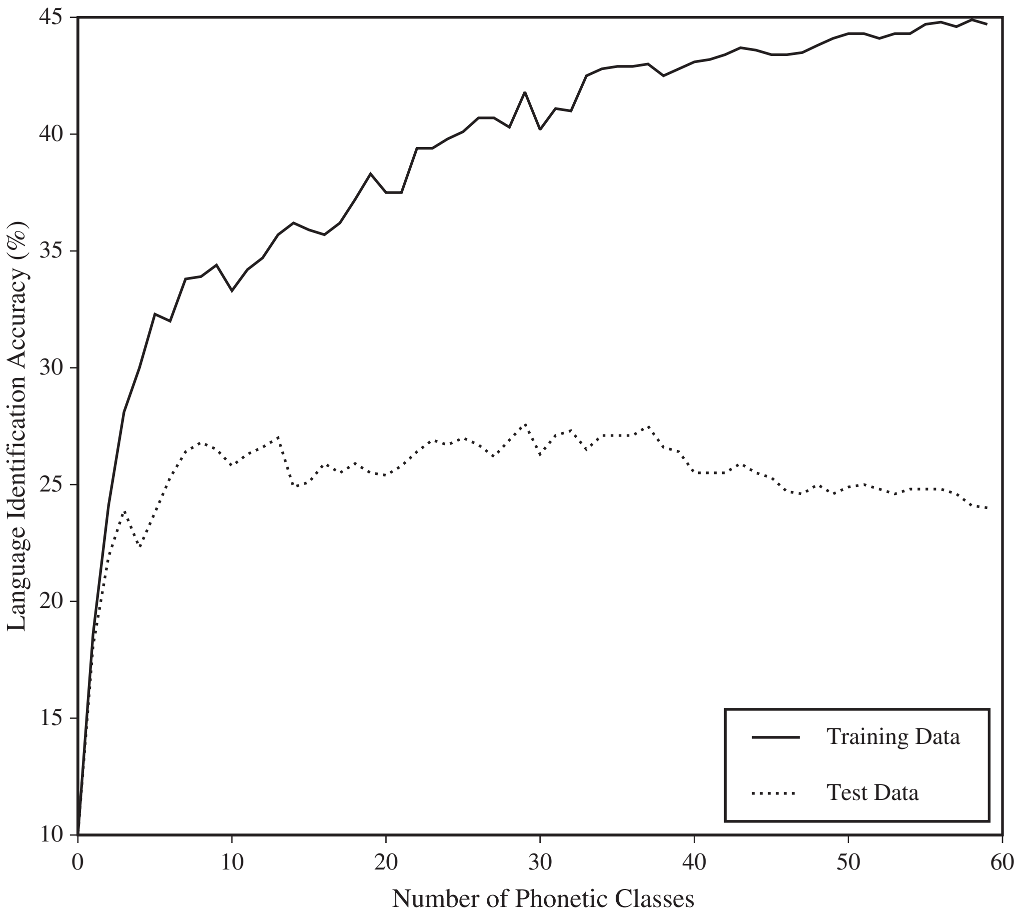

Figure 3.16: Language identification accuracy of segment duration model as the number of phonetic classes is varied

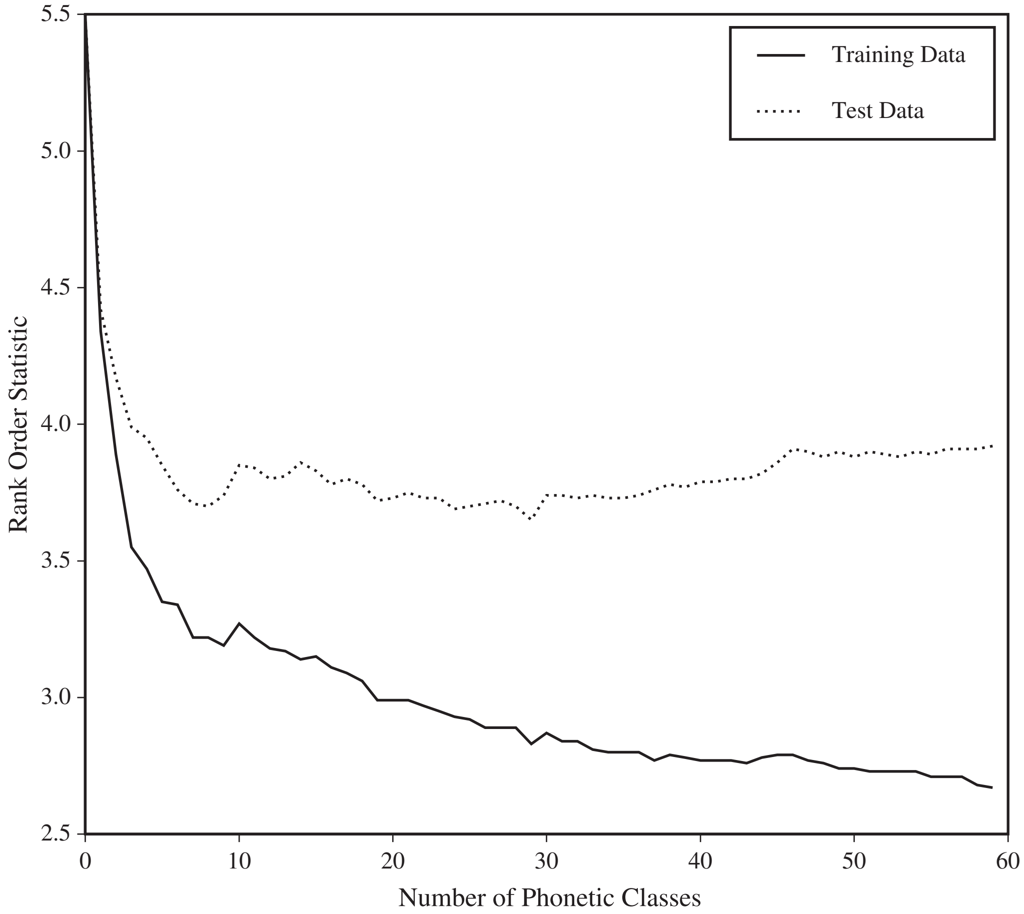

Figure 3.17: Rank order statistic of segment duration model as the number of phonetic classes is varied

Using the above assumptions, continuous probability density functions which model the segment-based acoustic feature vectors for each phonetic class in each language can be used for the acoustic model. The acoustic feature vectors in this case contain the values of each of the 14 MFCC's and 14 delta MFCC's averaged over the length of each segment. For this thesis, the acoustic feature vectors are modeled with mixtures of diagonal Gaussian density functions. To insure proper amounts of training data for each mixture of Gaussians, the number of Gaussians used to model each phonetic class follows the equation

$$n_{gaussians} = \begin{cases} n_{max} & \text{if } k/100 > n_{max} \\ \lceil k/100 \rceil & \text{otherwise} \end{cases} \tag{3.20}$$

where $n_{gaussians}$ is the number of Gaussians used in the mixture Gaussian model of a particular phonetic class for a particular language, n_{max} is the maximum number of Gaussians allowed in each mixture, and k is the number of training vectors for the phonetic class in that particular language.

To find an adequate maximum number of Gaussians to use within each mixture, the performance of the acoustic model was examined as the maximum number of Gaussians was varied from 1 to 28. Figures 3.18 and 3.19 shows the performance of the acoustic model over varying numbers of Gaussians per mixture. As can be seen in Figure 3.19, the performance of the acoustic model begins to level off as the maximum number of Gaussians is increased beyond 13. Using a maximum of 16 Gaussians per mixture, the acoustic model achieves a language identification accuracy of 37.9% with a rank order statistic of 3.27.

3.5.6 System Integration

To complete the ALI system, each of the individual models must be integrated into one system. Since the system is seeking the language which is most likely given the acoustic information, the probability scores from each individual model for an utterance must be combined to provide one probability score for each language. Using the probabilistic framework, this can be accomplished with the following expression:

$$\max_i \Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i) \Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) \Pr(\hat{C} \mid L_i). \tag{3.21}$$

To prevent underflow errors in the computation, the logarithm of the expression can be taken to yield the following expression:

$$\max_i \log \left(\Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i) \Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) \Pr(\hat{C} \mid L_i) \right). \tag{3.22}$$

The expression in (3.22) can further be expressed as a sum of logarithms yielding

$$\max_i \left(\log(\Pr(\vec{\mathbf{a}} \mid \hat{C}, \hat{S}, \vec{\mathbf{f}}, L_i) + \log \Pr(\hat{S}, \vec{\mathbf{f}} \mid \hat{C}, L_i) + \log \Pr(\hat{C} \mid L_i) \right). \tag{3.23}$$

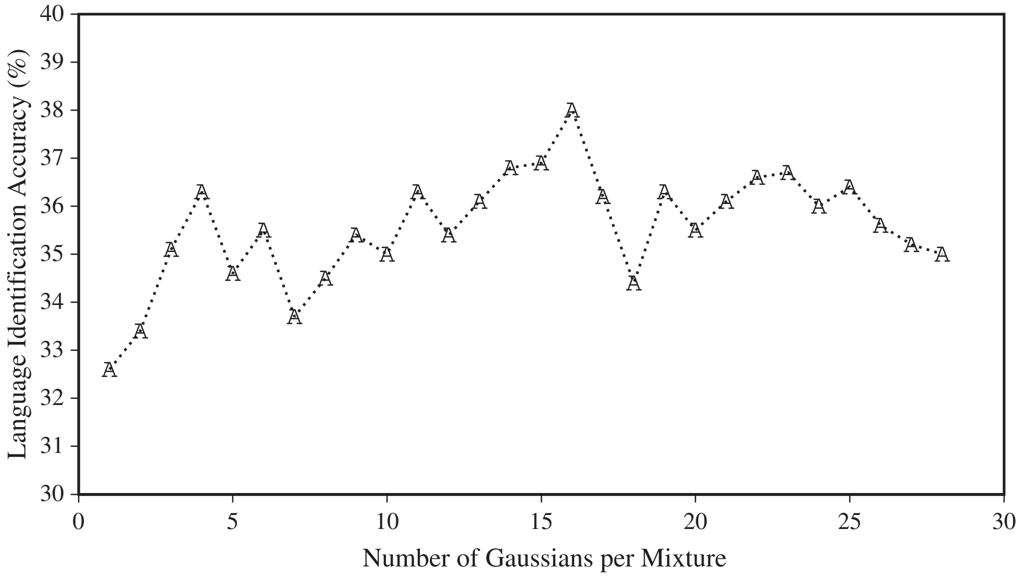

Figure 3.18: Accuracy of acoustic model as the number of Gaussians per mixture is varied

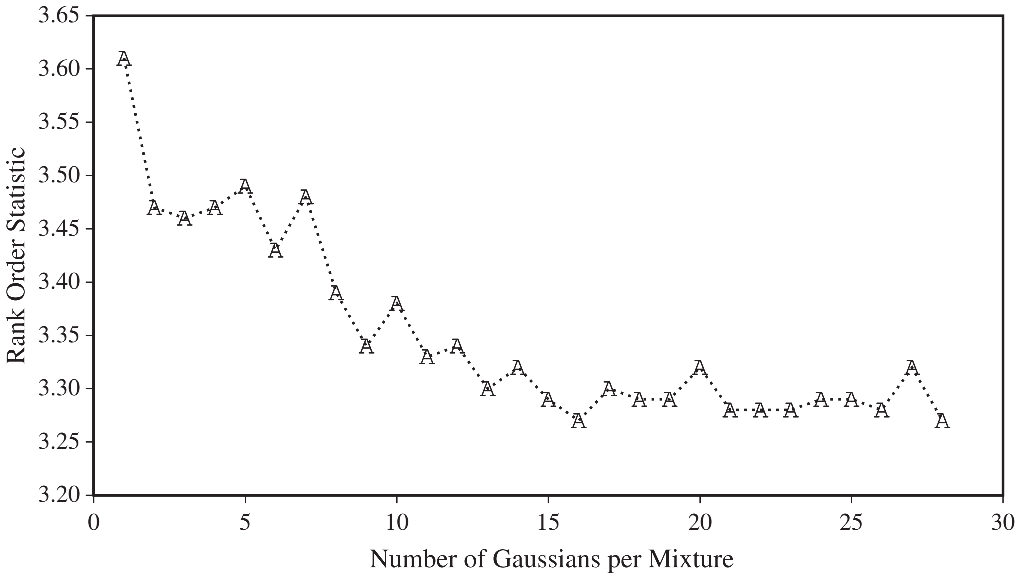

Figure 3.19: Rank order statistic of acoustic model as the number of Gaussians per mixture is varied

Model Name	Model Description
Language Model	Interpolated trigram models using 59 phonetic classes
Segment Duration Model	Non-parametric probability distributions using 29 phonetic classes
F0	Mixtures of 9 full covariance Gaussians
Acoustic Model	Mixtures of 16 diagonal Gaussians using 59 phonetic classes

Table 3.5: Summary of individual models used in final ALI system

Thus, the log likelihood score for each language can simply be represented as the sum of the log likelihood scores for each of the individual models.

Using the log likelihood approach, the language, acoustic, F0, and segment duration models can easily be integrated into the final system. A summary of the individual models that are used in the final system is presented in Table 3.5. The system uses 59 phonetic classes in the representation of the phonetic string $\hat{C}$. However, because the segment duration model cannot be sufficiently trained when it uses 59 classes, the elements of $\hat{C}$ are collapsed into 29 phonetic classes for the segment duration model.

Unfortunately, when the final system uses the simple addition of log likelihood scores with equal weights as described above, the final log likelihood score for each language is dominated by the F0 model score. To examine the scores of each model in more detail, the final log likelihood scores for each model for each utterance can be converted into the *a posteriori* language probabilities. For example, the language model can be represented with the following equation:

$$\Pr(\hat{C} \mid L_i) = \frac{\Pr(L_i \mid \hat{C}) \Pr(\hat{C})}{\Pr(L_i)}. \tag{3.24}$$

From (3.24) the *a posteriori* language probability can be expressed as

$$\Pr(L_i \mid \hat{C}) = \frac{\Pr(\hat{C} \mid L_i) \Pr(L_i)}{\Pr(\hat{C})} = k \Pr(\hat{C} \mid L_i) \tag{3.25}$$

where k is a constant value for each utterance. The value of k can be calculated easily given the condition that the *a posteriori* language probabilities over all i must sum to one. Once k is calculated for a specific utterance, the language model scores for that utterance can easily be converted into the *a posteriori* language probabilities. The *a posteriori* probabilities for any of the other models can be found in the same fashion.

Model Name	Average *A Posteriori* Probability of Top Choice	Actual Language ID Accuracy
Language Model	.769	.417
Segment Duration Model	.470	.282
F0 Model	.946	.207
Acoustic Model	.970	.338

Table 3.6: Average *a posteriori* probability of top choice language vs. actual language identification accuracy for each model on data jackknifed from the training set

The average *a posteriori* probability of a model's top choice language provides a measure of the certainty in which a model believes its top choices are correct. Therefore, it is expected that a sound model will achieve an actual accuracy which is approximately equal to the average *a posteriori* probability of its top choice. The average *a posteriori* language probabilities of the top choice language for each model was found from utterances which were jackknifed from the training data.[8] Table 3.6 shows the comparison between the average *a posteriori* language probability of the top choice language and the actual language identification accuracy for each model. As can be seen in the table, the average top choice probability is larger than the actual language identification accuracy for each of the models. This indicates that the top choice probabilities are being inflated significantly higher than they actually should be. This may be due to that fact that the assumptions regarding the independence of segments (or frames) which are made in each of the models allow biases due to speaker and channel dependencies to accumulate across the length of the utterance. This effect is most prevalent in the F0 and acoustic models but is also present to a lesser degree in the language and segment duration models.

To compensate for the discrepancy between the average top choice probability and the language identification accuracy of each model, the log likelihood score of each model can be multiplied by an artificial scaling factor. Multiplying the log likelihood scores of a model by a scaling factor will effectively compress or expand the range of *a posteriori* probabilities for that model. Scaling factors less than one will compress the range of the *a posteriori* probabilities causing the set of probabilities to become more uniform. For the four models used in the system, scaling factors which compress the range of *a posteriori* probabilities are appropriate. More specifically, the scaling factor

[8]The training data was divided into 5 unique sets of 40 speakers per language for training and 10 speakers per language for jackknifed testing for this experiment.

Model Name	Log Likelihood Scaling Factor
Language Model	.2375
Segment Duration Model	.4250
F0	.0108
Acoustic Model	.0174

Table 3.7: Log likelihood scaling factors for each model

for each model is chosen to adjust the top choice average *a posteriori* probability of the model so it is equal to the model's actual language identification accuracy. The scaling factors for each model are shown in Table 3.7. Using the scaling factors shown in Table 3.7, the final system was able to achieve a language identification accuracy of 48.6% with a rank order statistic of 2.51.

Chapter 4

Analysis

4.1 Overview

When evaluating an ALI system it is important to examine the system's performance from a variety of perspectives. Examining a simple statistic such as the system's overall language identification accuracy or rank order statistic may not provide sufficient insight into the various factors which contribute to the system's performance. It is important to understand how the performance is affected as various test conditions are altered. It is also important to understand the types of errors that are made and the severity of these errors. Some of the important issues that should be examined can be summarized in the following questions:

- Which types of information (i.e., phonetic, acoustic, prosodic) are most useful for language identification?
- How is the system performance affected by the length of an utterance?
- How is the system performance affected by the vocabulary and speaking style constraints of an utterance?
- How is the system performance affected by the size of the training set?
- What types of errors does the system make?
- How is the system performance affected by alterations in the language set?
- What are the receiver-operator characteristics of the system?

An examination of these issues is important in determining the strengths and weaknesses of a system. A clear understanding of a system's advantages and drawbacks is necessary if efforts to improve the system's design and performance are to be successful.

4.2 Performance of Individual Models

In examining the performance of the ALI system, it is desirable to understand which information utilized by the system is the most useful. To accomplish this, the performance of each of the different components of the system can be examined separately as well in combination with other components. As a review, Table 4.1 summarizes the the properties of the different models used by the system.

The performance of the ALI system using different combinations of the system's models is shown in Table 4.2. The results throughout the table show that the language model is the most important model for language identification. This finding supports House and Neuburg's belief that the phonotactic constraints of languages provide information that is useful for language identification. However, the results also show that improvements in performance can be gained by using additional information to supplement the language model.

The table shows how each of the other models (i.e., the acoustic, duration and F0 models) performs when used in conjunction with the language model. Despite the fact that the F0 model is the weakest of all of the models when used on an individual basis, the F0 model contributes more than the acoustic or duration models when used in conjunction with the language model. Similarly, when the F0 and duration models are combined to form the prosodic model, the prosodic model contributes more to the overall system than the acoustic model. One possible explanation for this behavior is that there may be a larger correlation between the information carried in the language and acoustic models than there is between the information of the language and prosodic models. As such, the prosodic model may be supplementing the language model with more independent information than the acoustic model.

Additionally, it is interesting to note that the performance of the system using only the combination of the prosodic and acoustic models is nearly as high as the

Model Name	Model Description
Language Model	Interpolated trigram models using 59 phonetic classes
Segment Duration Model	Non-parametric probability distributions using 29 phonetic classes
F0	Mixtures of 9 full covariance Gaussians
Acoustic Model	Mixtures of 16 diagonal Gaussians using 59 phonetic classes

Table 4.1: Summary of individual models used in final ALI system

performance using only the language model. This further indicates that, while the phonotactic constraints of languages may be powerful, prosodic and acoustic information are also useful for language identification.

Set of Models	Language Identification Accuracy	Rank Order Statistic
Language Model	41.7%	2.78
Acoustic Model	37.9%	3.27
Duration Model	27.6%	3.65
F0 Model	23.1%	4.01
Duration + F0 (i.e., Prosodic Model)	32.7%	3.27
Language + F0	45.8%	2.61
Language + Acoustic	45.1%	2.69
Language + Duration	42.8%	2.72
Language + F0 + Acoustic	47.6%	2.60
Language + F0 + Duration	46.2%	2.57
Language + Acoustic + Duration	45.5%	2.63
Language + Prosodic	46.2%	2.57
Language + Acoustic	45.1%	2.69
Prosodic + Acoustic	41.0%	2.86
Complete System	48.6%	2.51

Table 4.2: System performance using varying sets of models

4.3 Performance Over Varying Utterance Lengths

The performance of the system as the test utterance length is varied is shown in Figures 4.1 and 4.2. These plots were obtained by examining the system performance using only the first t seconds of each utterance where t was varied from 1 second to 45 seconds. For each value of t only the utterances with a length greater than t are used. As expected, the system performs better as the utterance length is increased. On the unconstrained utterances, the system improved from an accuracy of 33.1% using 2 seconds of speech to 47.0% using 10 seconds to 56.8% using 45 seconds.

Figure 4.3 shows the performance of the individual models over time. Figure 4.3 reveals that the language model's performance has a larger increase as the utterance length is increased than any of the other models. In fact, the language model performs

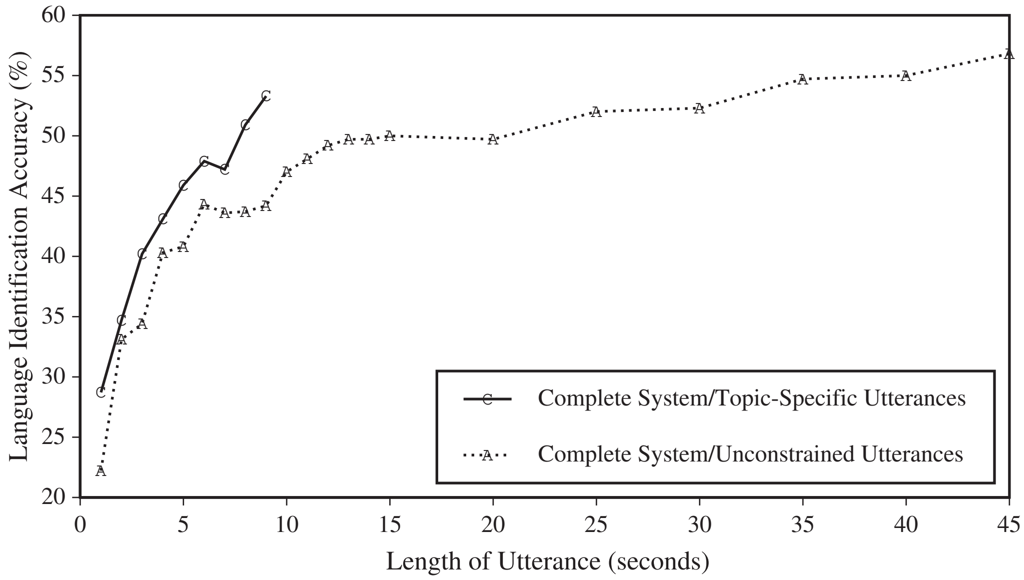

Figure 4.1: Language identification accuracy over varying test utterance length

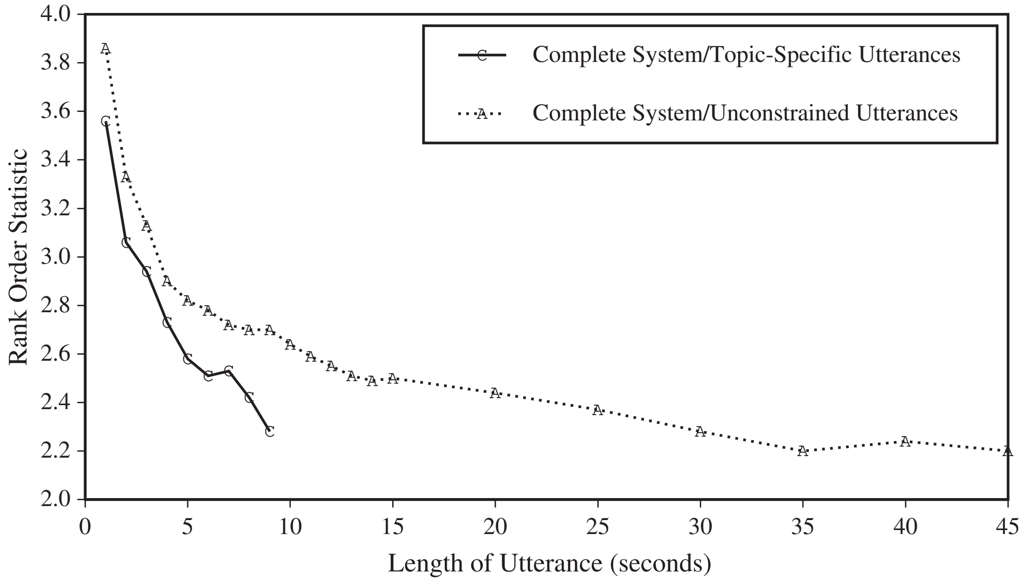

Figure 4.2: Rank order statistic over varying test utterance length

as well as the complete integrated system as the length is increased past 40 seconds. The F0 and duration models also incur significant improvements in their performance as the utterance length increases although not as large as the improvement in the language model. The acoustic model has a considerably smaller improvement in performance as the length is increased, and in fact, has no improvement as it is increased beyond 15 seconds. This may indicate that little additional information is gained by the acoustic model from any more than a few observations of each phone.

It is interesting to note that for utterance lengths of 3 seconds or less, the acoustic model outperforms all of the other models. However, as the utterance length is increased beyond 20 seconds, both the language and duration models outperform the acoustic model. These results indicate that a language identification strategy which reduces the weight placed on the acoustic model score and increases the weight placed on the language model score as the test utterance length becomes longer may be more appropriate than the static weighting system that was used in this thesis.

4.4 Performance Using Utterance Constraints

Figures 4.1 and 4.2 also show the performance of the system using two different utterance constraints. The figures show the difference in performance using the topic-specific utterances versus the unconstrained utterances. As can be seen, the system performed significantly better using the topic specific utterances. Using nine seconds of speech from test utterances, the system achieved an accuracy of 53.3% on the topic-specific utterances while only achieving 44.2% on the unconstrained utterances. Figures 4.4, 4.5, 4.6, and 4.7 show the performance of each of the individual models on the topic-specific and unconstrained utterances. The figures show that each of models (with the possible exception of the duration model) performs better on the topic-specific utterances.

Because of the vocabulary constraints of the topic-specific utterances, it is expected that the language model component of the system would perform better on these utterances than on the unconstrained utterances. However, the figures also reveal that the language identification performance of the acoustic and prosodic models was also better using the topic-specific utterances. This may partially be due to the fact that some acoustic and prosodic information, such as the stress patterns of specific words, is correlated with the vocabulary. However, the acoustics and prosodics may also be affected by fundamental differences in the speaking styles used in the topic-specific and unconstrained utterances. The topic-specific utterances were all spontaneous replies to queries while the unconstrained utterances were not limited in any fashion. In fact, the unconstrained utterances contained examples of both spontaneous and read speech, which are known to be different in their prosodic nature [3].

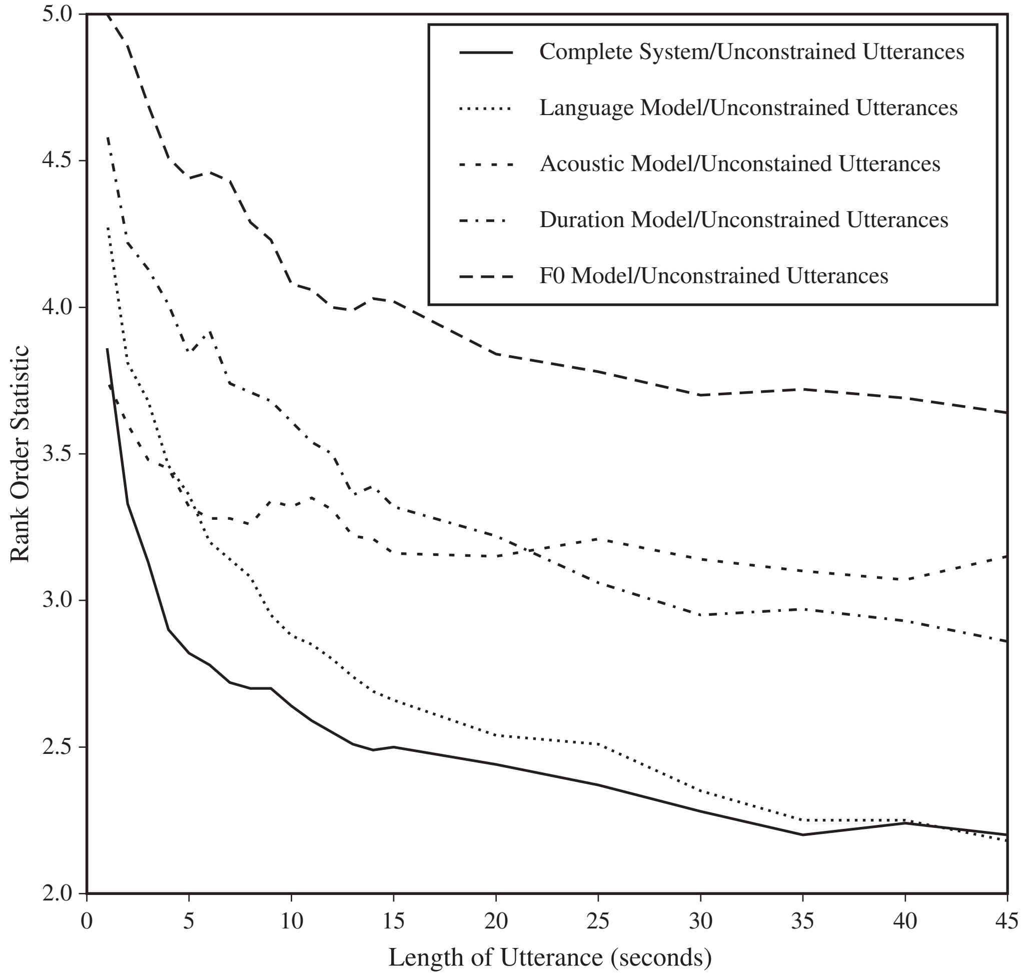

Figure 4.3: Performance of individual models over varying test utterance length

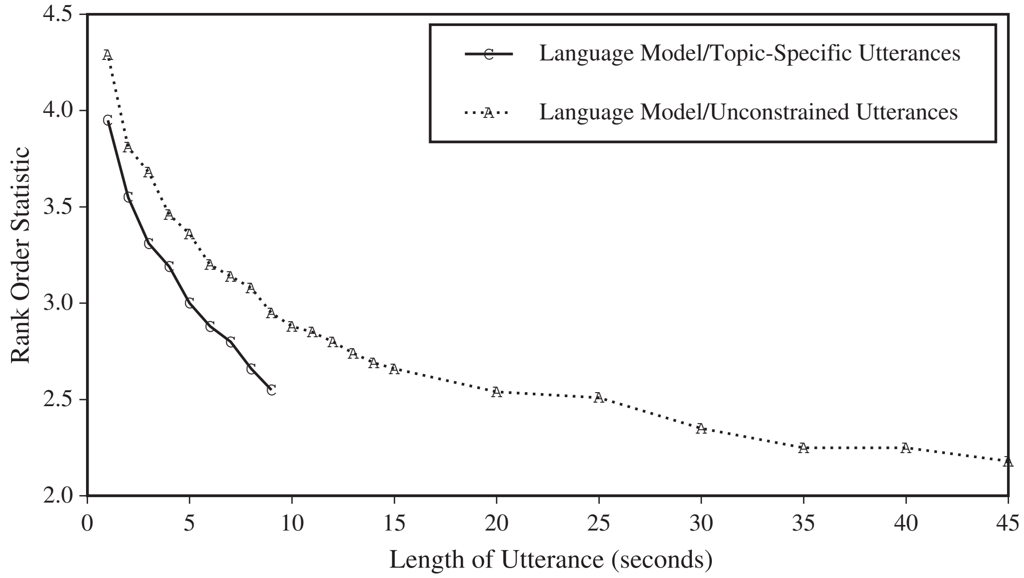

Figure 4.4: Language model performance: topic-specific vs. unconstrained utterances

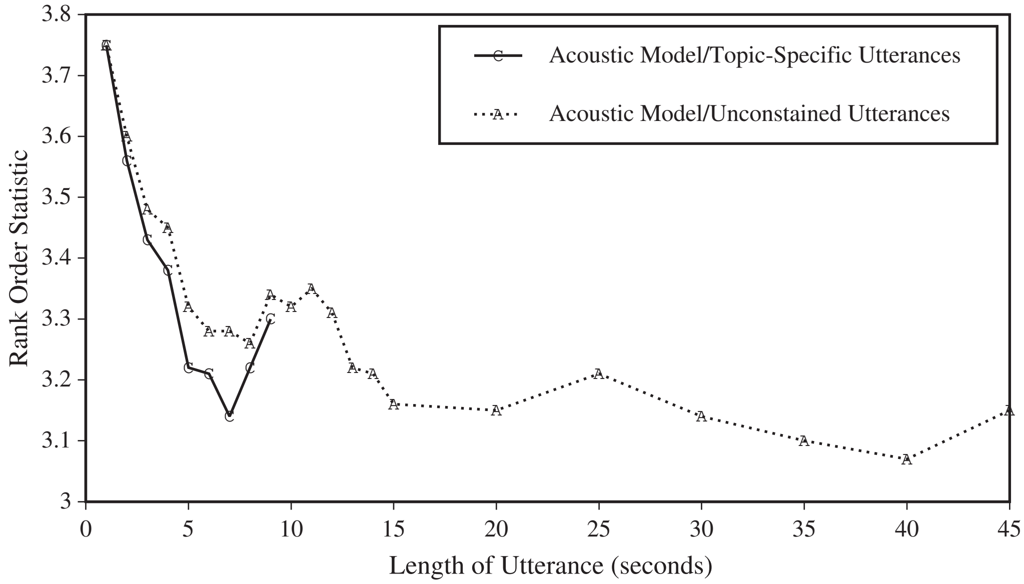

Figure 4.5: Acoustic model performance: topic-specific vs. unconstrained utterances

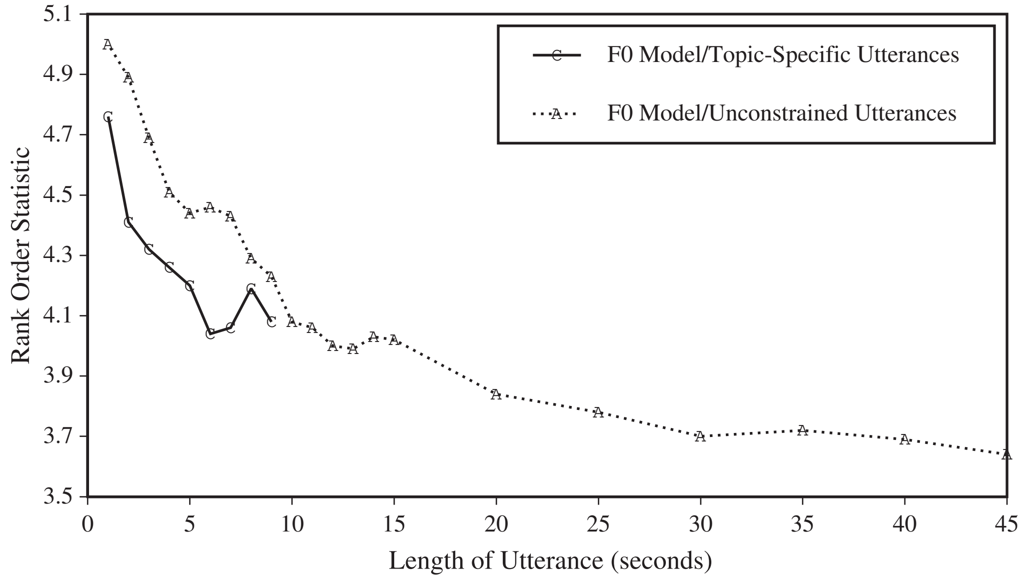

Figure 4.6: F0 model performance: topic-specific vs. unconstrained utterances

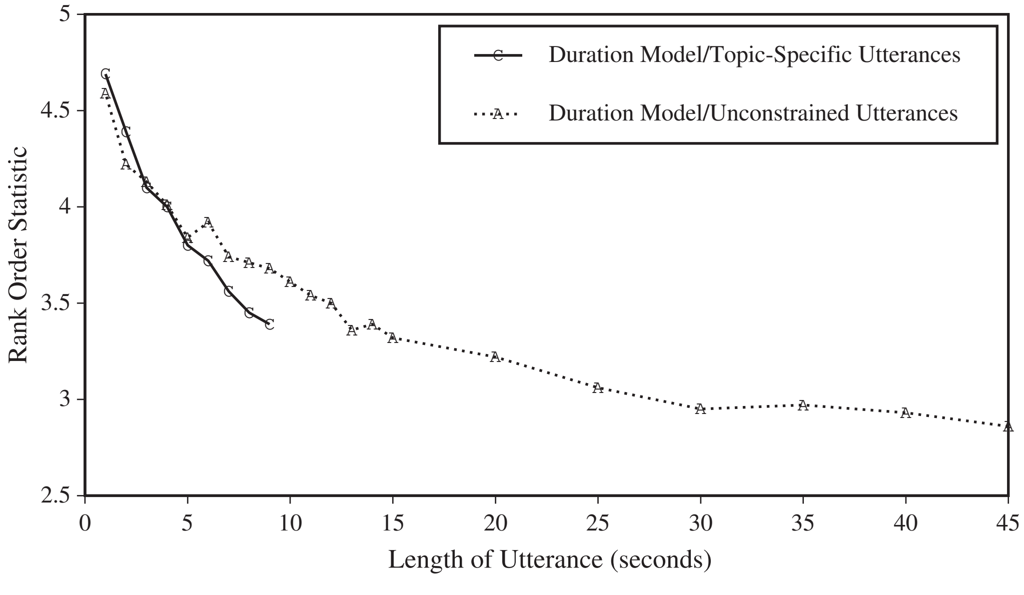

Figure 4.7: Duration model performance: topic-specific vs. unconstrained utterances

4.5 Performance Over Varying Training Set Sizes

Figure 4.8 shows the overall performance of the system as the training set size is varied. While the performance on the two data sets is converging as more training speakers are used, there is still a large gap between the performance on training and testing data even when the training set size is increased to 50 speakers per language. This indicates that there is still plenty of room for improvement in the system.

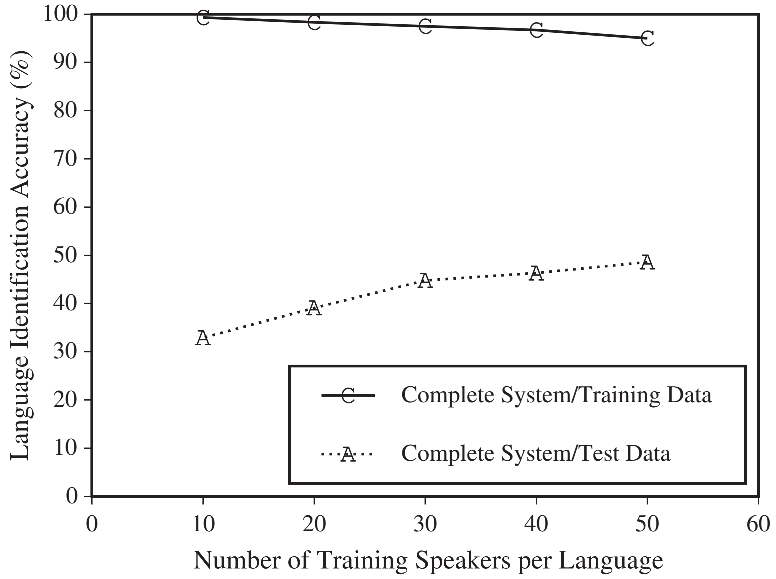

Figure 4.8: System performance over varying training set sizes

Figures 4.9, 4.10, 4.11 and 4.12 show the performance over varying training set sizes for each of the individual models. These figures show that the language and acoustic models have a significant gap between the training and testing performance. This is most likely due to the fact that the language and acoustic models contain far more parameters to be trained than the duration and F0 models. There is also a significant (albeit smaller) gap between the training and test set performance of the duration model. It is possible that the large gaps between the training and test set performance in the language, acoustic, and duration models could be decreased if the error rate of the phonetic recognizer was decreased. The F0 model, on the other hand, has near convergence of its training and test set performance with a training set of 50 speakers per language. This would indicate that the F0 model could withstand a significant increase in its complexity without suffering from a lack of training data.

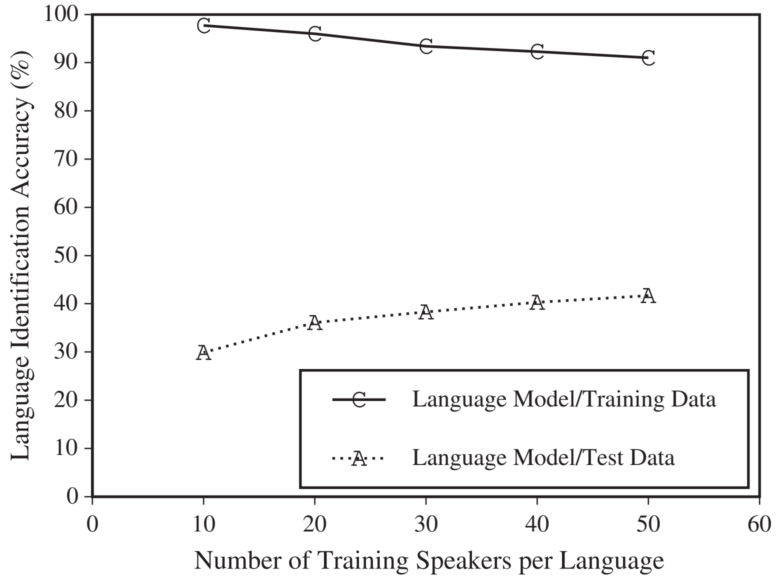

Figure 4.9: Language model performance over varying training set sizes

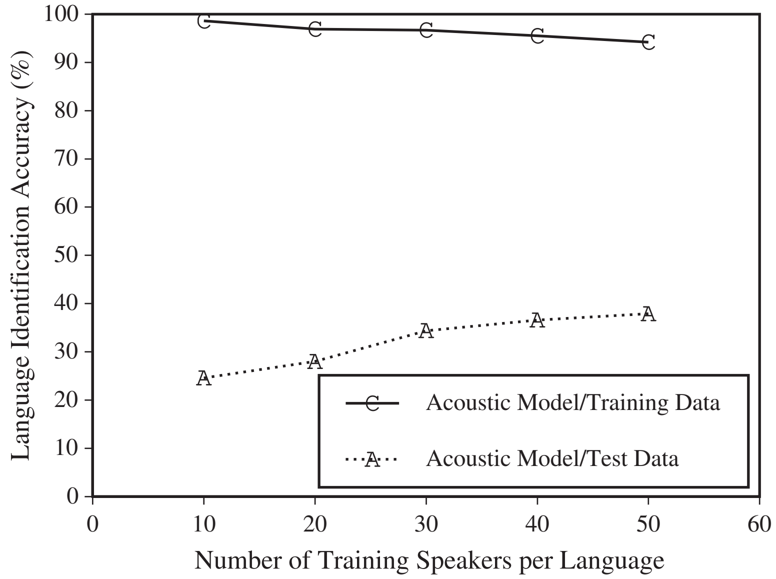

Figure 4.10: Acoustic model performance over varying training set sizes

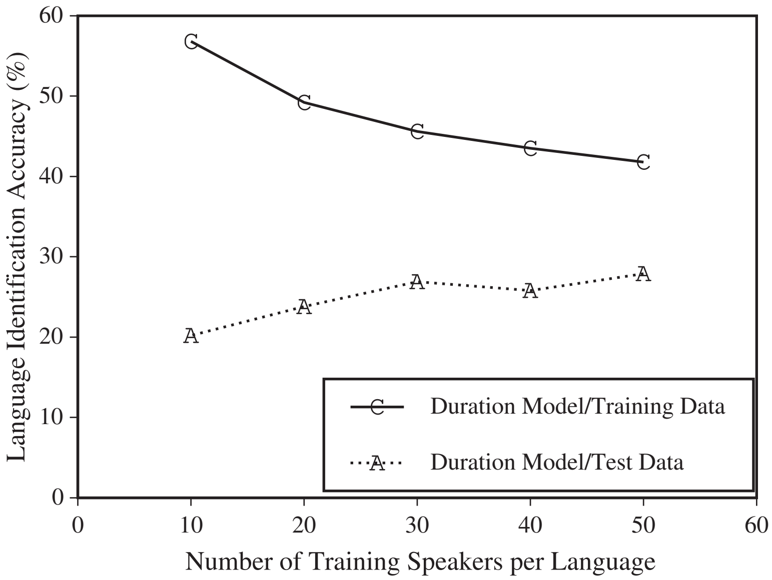

Figure 4.11: Duration model performance over varying training set sizes

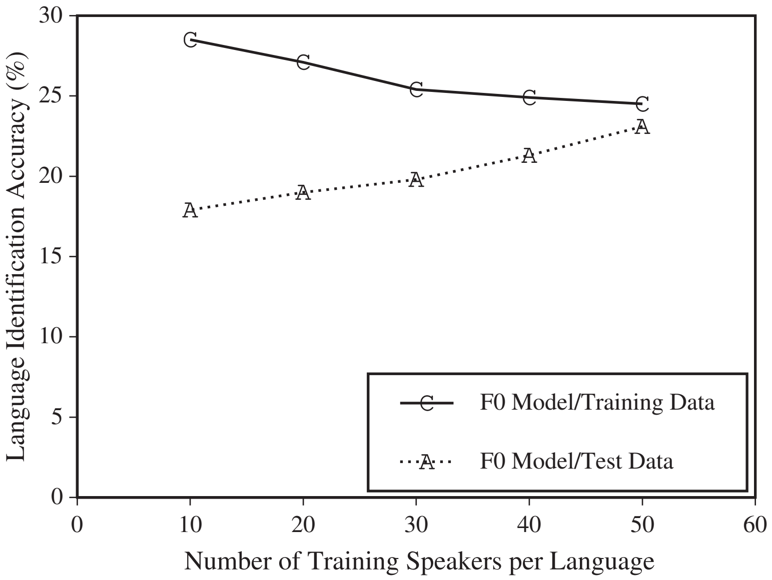

Figure 4.12: F0 model performance over varying training set sizes

4.6 Analysis of Confusions

The confusion matrix for the complete system is shown in Table 4.3. Several important observations can be made about the errors present in the confusion matrix. First, there appears to be a bias in the system towards choosing an Indo-European language (English, German French, Spanish or Farsi) as the system's top choice. Although only 51% of the test utterances are from an Indo-European language, 61% of the test utterances are classified by the system as one of the five Indo-European languages. Table 4.4 shows a breakdown of the confusions which occur between Indo-European and non-Indo-European languages. As a second observation, there are a few pairs of languages which have significantly larger than average confusion rates. In particular, the pairs English-German and German-French appear very confusable. This is understandable given that these languages are all from the Indo-European language family. However, Japanese and French, which do not belong to same family, are also very confusable.

Input	Output Hypothesis									
Utterance	Eng	Ger	Fre	Spa	Far	Tam	Vie	Man	Kor	Jap
English	50.4	8.7	5.2	11.3	5.2	2.6	6.1	4.3	4.3	1.7
German	27.1	43.2	16.1	4.2	5.1	0.8	0.8	0.8	0.0	1.7
French	11.3	7.8	63.5	2.6	2.6	0.0	0.9	3.5	3.5	4.3
Spanish	7.2	8.1	9.0	50.5	6.3	5.4	2.7	0.9	2.7	7.2
Farsi	3.6	16.2	9.9	2.7	46.8	0.9	7.2	2.7	6.3	3.6
Tamil	1.8	1.8	2.7	14.2	8.8	51.3	6.2	1.8	4.4	7.1
Vietnamese	9.3	3.7	7.5	5.6	8.4	6.5	38.3	3.7	10.3	6.5
Mandarin	2.8	16.5	1.8	1.8	6.4	0.9	0.9	58.7	2.8	7.3
Korean	8.3	6.4	12.8	7.3	8.3	0.0	4.6	3.7	42.2	6.4
Japanese	2.8	9.8	18.8	8.9	4.5	0.9	4.5	5.4	4.5	40.2

Table 4.3: Confusion matrix of complete system (all values are percentages)

4.7 Performance Over Varying Language Sets

Examining the performance of the system on the task of pairwise language identification may provide a clearer picture of the confusions that can occur amongst the 10 different languages. Table 4.5 shows the performance of the system when the language set is limited to a pair of languages. All 45 combinations of language pairs were

Input Utterance	Output Hypothesis Indo-European	Other
Indo-European	85.1	14.9
Other	36.2	63.8

Table 4.4: Confusion matrix of Indo-European vs. non-Indo-European languages

tested and are shown in the table. The average performance of each language within the pairwise tests is also shown. For example, in the context of English-L_i where L_i can be any particular language other than English, the system had an average performance of 78.7% accuracy across all L_i.

The language with the largest average performance in the pairwise tests was Tamil which achieved an average accuracy of 88.4% in the Tamil-L_i pairwise tests. The best pairwise performance of the entire system was 92.6% for the Tamil-German pair. These results indicate that Tamil is the language which is most dissimilar from the rest of the languages based on the information used in the ALI modeling.

As might be expected, the language pair with the poorest pairwise performance was the English-German pair with an accuracy of only 63.5%. The French-German pair was also highly confusable with an accuracy of only 75.5%. In fact, the average performance across all pairs of the four European languages (English, German, French and Spanish) was only 75.7% which is considerably lower than the average of 83.2% across all language pairs. The system also experienced low pairwise performances with the Japanese-French pair (76.6%) and the Korean-English pair (74.1%). The confusions between these pairs are difficult to explain since both pairs contain languages from different language families.

To further demonstrate the importance of the particular set of languages on the performance of the system, several experiments were conducted using three different sets of 5 languages. Table 4.6 shows the confusion matrix and performance of the system using the five Indo-European languages. Table 4.7 shows the confusion matrix and performance of the system using the five non-Indo-European languages. Table 4.8 shows the confusion matrix and performance of the system using a set of five diverse languages. As can be seen, the system performs much better on the two sets which contain languages from different language families than it does on the set containing languages that are all from the Indo-European family.

To attempt to extract any hidden structure that may be present in the pairwise performance matrix in Table 4.5, hierarchical clustering can be performed using the separate rows of the matrix. By filling in a value of 50 for all of the diagonal elements

	Eng	Ger	Fre	Spa	Far	Tam	Vie	Man	Kor	Jap	Avg
Eng	-	63.5	77.4	77.9	77.0	86.4	81.1	83.9	74.1	87.2	78.7
Ger	63.5	-	75.5	79.5	77.7	92.6	91.1	85.4	88.1	84.3	82.0
Fre	77.4	75.5	-	80.5	87.6	92.5	87.4	90.6	82.6	76.6	83.4
Spa	77.9	79.5	80.5	-	86.9	78.6	83.9	87.7	83.6	78.9	81.9
Far	77.0	77.7	87.6	86.9	-	90.6	78.4	82.7	75.0	84.3	82.2
Tam	86.4	92.6	92.5	78.6	90.6	-	85.5	91.9	88.3	89.3	88.4
Vie	81.8	91.1	87.4	83.9	78.4	85.5	-	86.1	80.6	84.5	84.4
Man	83.9	85.4	90.6	87.7	82.7	91.9	86.1	-	84.4	82.8	86.2
Kor	74.1	88.1	82.6	83.6	75.0	88.3	80.6	84.4	-	80.5	81.9
Jap	87.2	84.3	76.6	78.9	84.3	89.3	84.5	82.8	80.5	-	83.2
Average Performance Across All Language Pairs:											83.2

Table 4.5: Performance of system on pairs of languages (all values are language identification accuracies in percentages)

Input Utterance	Output Hypothesis				
	Eng	Ger	Fre	Spa	Far
English	52.2	15.7	7.0	15.7	9.6
German	27.1	44.9	16.1	6.8	5.1
French	13.0	11.3	67.0	5.2	3.5
Spanish	8.1	15.3	12.6	55.9	8.1
Farsi	3.6	24.3	9.9	4.5	57.7
Overall System Accuracy: 55.4					

Table 4.6: Confusion matrix and performance of system using the 5 Indo-European languages

Input Utterance	Output Hypothesis				
	Tam	Vie	Man	Kor	Jap
Tamil	70.8	6.2	5.3	8.8	8.8
Vietnamese	8.4	52.3	8.4	17.8	13.1
Mandarin	2.8	1.8	76.1	7.3	11.9
Korean	0.9	6.4	11.9	63.3	17.4
Japanese	3.5	4.5	15.2	8.0	68.8
Overall System Accuracy: 66.4					

Table 4.7: Confusion matrix and performance of system using the 5 non-Indo-European languages

Input Utterance	Output Hypothesis				
	Eng	Fre	Far	Tam	Man
English	61.7	12.2	11.3	3.5	11.3
French	20.0	67.8	5.2	0.9	6.1
Farsi	9.0	11.7	63.1	0.9	15.3
Tamil	8.0	4.4	10.6	72.6	4.4
Mandarin	6.4	2.8	11.0	2.8	77.1
Overall System Accuracy: 68.2					

Table 4.8: Confusion matrix and performance of system using 5 diverse languages

of the matrix, each row can be viewed as a vector.[1] Figure 4.13 shows the results of the hierarchical clustering using the Euclidean distance as the similarity measure for the vectors. As might be expected, the four European languages are all clustered together in one branch with the Germanic and Romance languages in separate sub-branches. Also clustered together in a separate branch of the tree are the only two tonal languages in the set, Mandarin and Vietnamese. However, the tree also contains the unexpected clusterings of Japanese with French and Korean with Farsi. These results offer confirmation that the system is capturing at least some of the fundamental differences which occur among languages and language families.

4.8 Receiver-Operator Characteristic

An examination of the receiver-operator characteristic (ROC) of the system can help determine the reliability of the system's scoring mechanism. The ROC reveals how a system's performance is affected as the certainty its top-choice score is varied. The ALI design in this thesis uses actual probability values to determine not only which language is the most likely candidate but also the certainty with which it believes its top choice to be correct. If the system is accurately determining the *a posteriori* language probability of its top-choice language candidate then the likelihood of the system's top-choice being correct will indeed increase as the *a posteriori* probability of the top-choice language candidate increases. The ROC of a system demonstrates how the system's performance is affected by the introduction of a *rejection* region. The ROC is calculated by setting a threshold on the system's top-choice score and *rejecting* all utterances which fall below that threshold. The threshold is varied to examine the system's performance as the rejection region is spanned from 0% rejection to 100% rejection. The standard ROC curve for the system is shown in Figure 4.14. The standard ROC curve shows the percentage of correctly identified utterances which are accepted (i.e., detection rate) against the percentage of incorrectly identified utterances which are accepted (i.e., false alarm rate) as the rejection region is varied. Because information about the absolute accuracy of the system is not readily apparent in the standard ROC curve, an alternate view of the ROC is shown in Figure 4.15. This figure shows how the system's overall accuracy is affected as the rejection region is increased. As can be seen in both figures, the system's performance does indeed improve as the utterances with the lowest scores are rejected. However, there is also plenty of room for improvement.

[1] In pairwise language identification, accuracies can be expected to range from a minimum of 50% to maximum of 100%. Thus, the more similar two languages are the more the system's accuracy should move towards 50%. Thus, a value of 50 is chosen to fill into the diagonal elements of the matrix since no language is more similar to any language than itself.

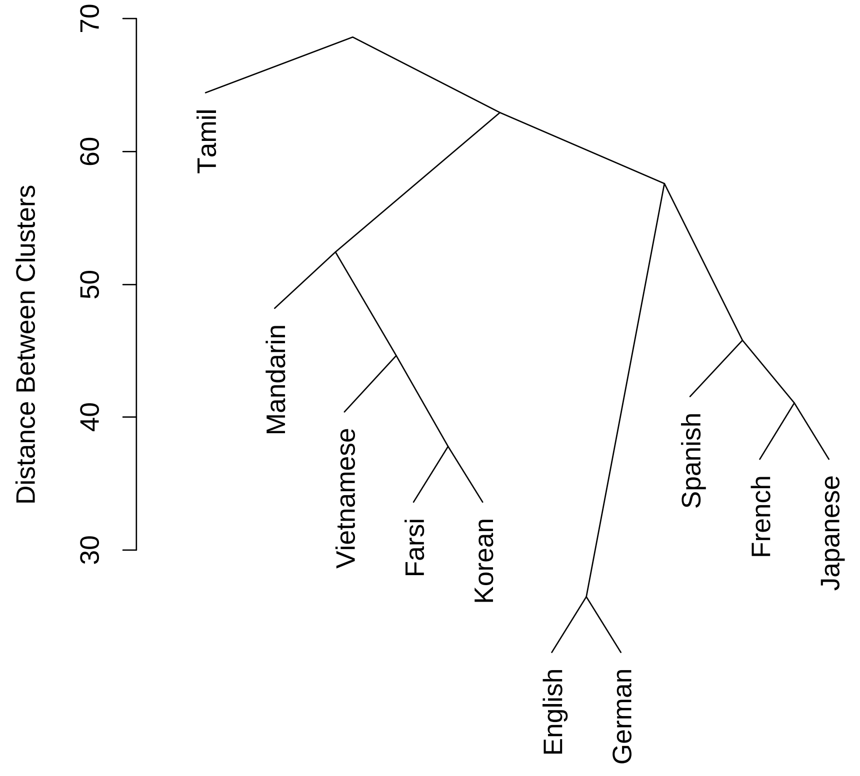

Figure 4.13: Clustering of languages based on performance in pairwise tests

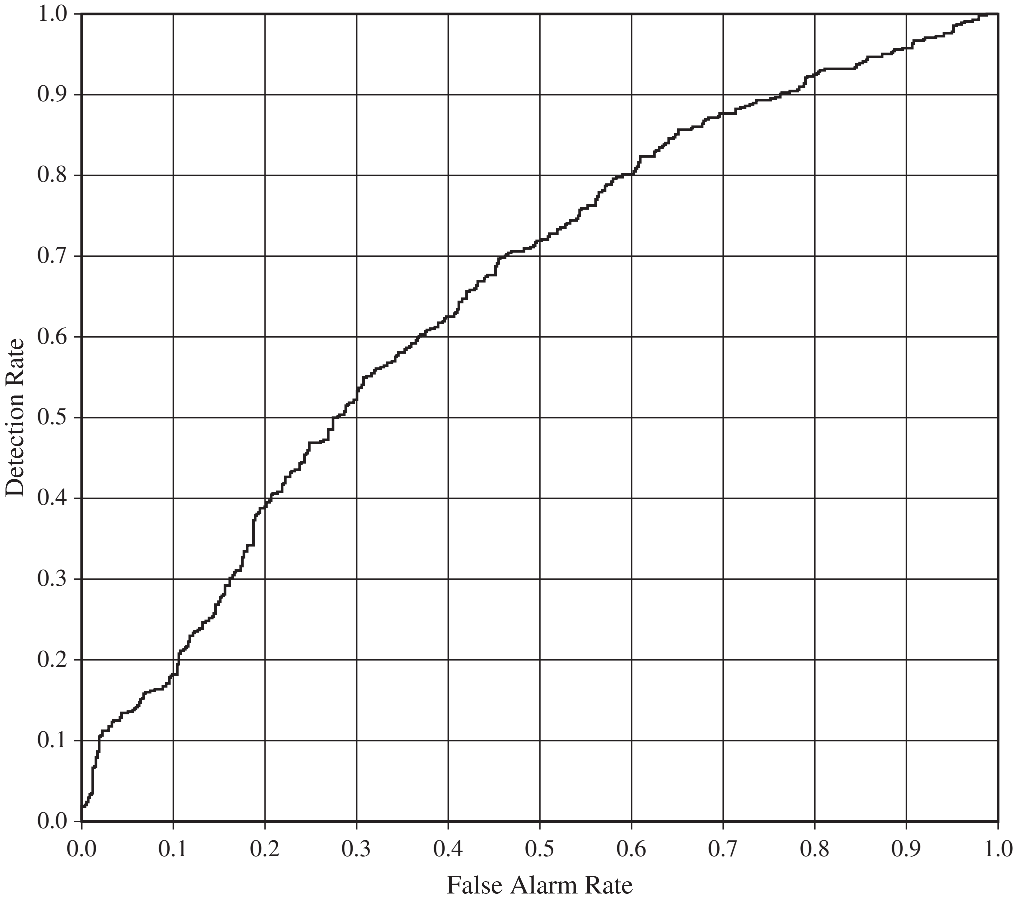

Figure 4.14: Standard ROC curve for the ALI system

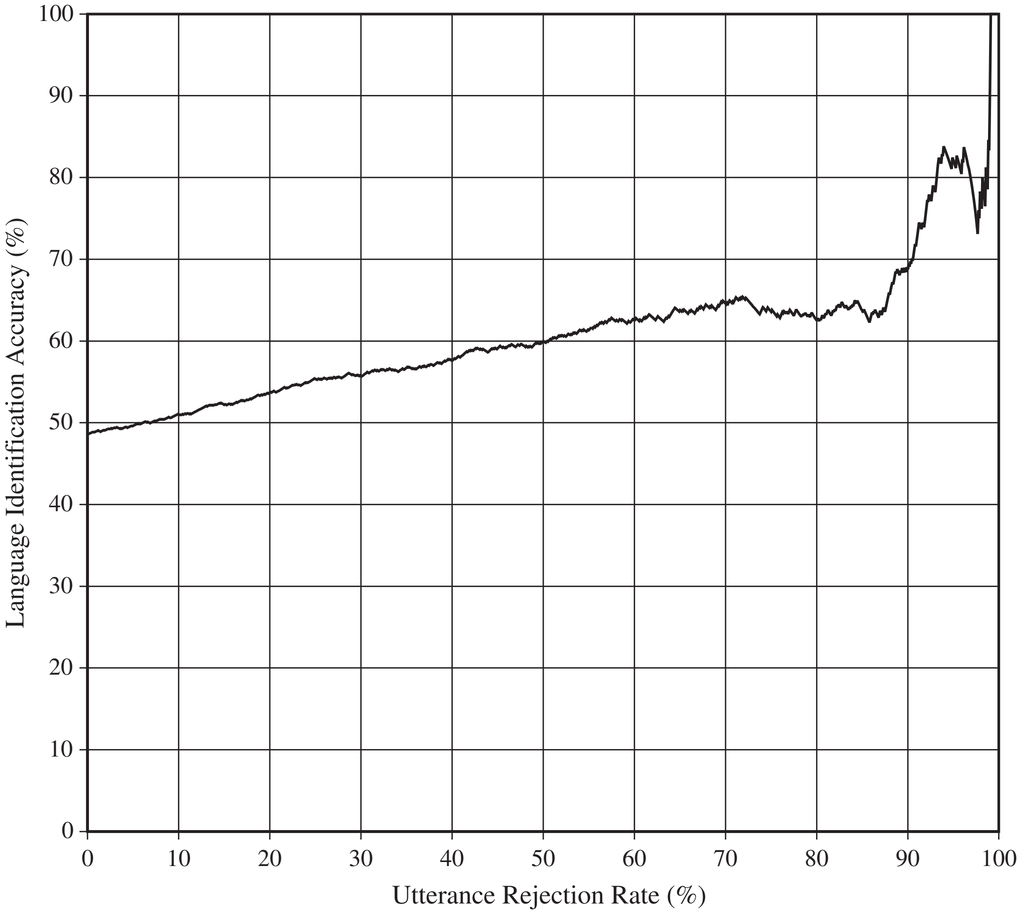

Figure 4.15: System accuracy over a varying rejection region

4.9 Rank Order Statistics

The rank order statistic can be useful in determining the severity of the errors that are incurred by the system. When a system fails to identify the correct language with its top choice, it is hoped that the correct language is at least the second or third choice of the system. Overall, the rank order statistic for the system is 2.51. Figure 4.16 shows how the system performs in the task of identifying the correct language of an utterance within the top n choices of its candidate list. As can be seen in the figure, the system identifies the correct language as one of its top three choices 76.3% of the time. Additionally, only 10.4% of the time is the correct language placed within the lower half of the candidate list. These results indicate that the system is able to provide a reliable list of alternative choices when its top choice is incorrect.

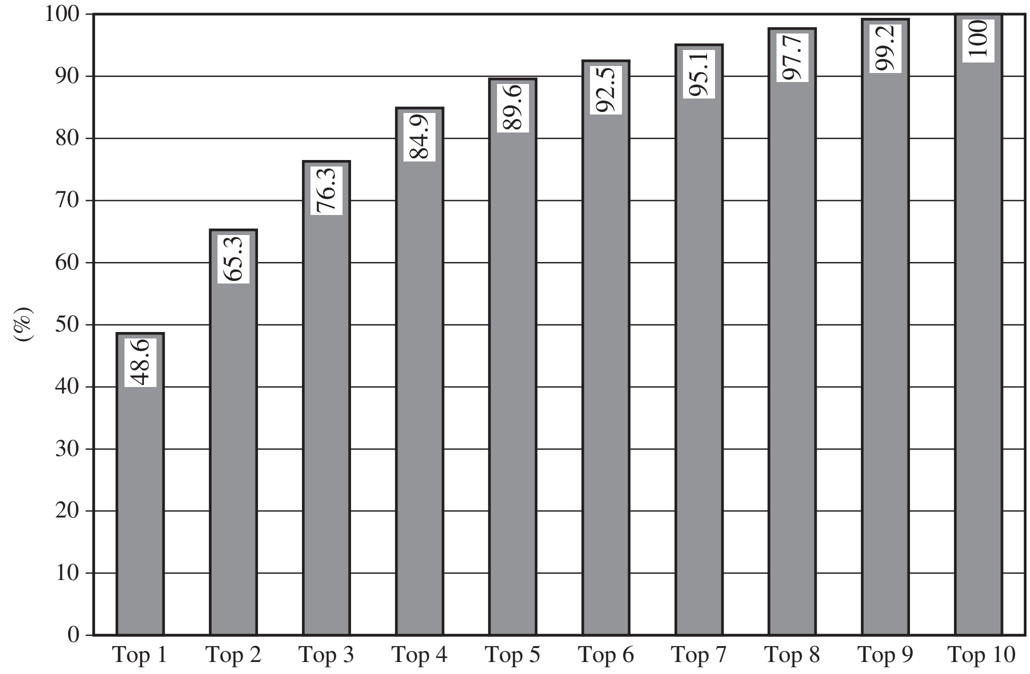

Figure 4.16: System accuracy in placing the correct language within the top n candidates

Chapter 5

Conclusion

5.1 Summary

This thesis has attempted to achieve three goals. The first goal was to present a formal probabilistic framework describing the ALI problem. This framework, which uses the ideas of House and Neuburg as a foundation, is presented in Chapter 2. The second goal was to present a new segment-based approach for ALI. This approach, which gains its structure from the probabilistic framework discussed in Chapter 2, is presented in Chapter 3. The third goal was to analyze and understand the various modeling decisions, assumptions, and test conditions which affect the system's performance. These analyses are presented in Chapters 3 and 4. Based on the investigation described in this thesis, we can draw several important, although tentative, conclusions. These are summarized below.

The House and Neuburg study indicated that the phonotactic constraints of languages are very strong and could prove extremely useful for ALI. The results of the experiments conducted for this thesis are supportive of this claim. The language model component of the ALI system, which was designed to capture the phonotactic constraints of the different languages, performed better than all of the other models combined. However, experiments also showed that House and Neuburg's proposal to represent the phonetic sequence with broad phonetic classes instead of detailed phonetic elements did not yield optimal performance. The results presented in Chapter 3 indicated that increasing the detail of the phonetic elements used to represent the phonetic sequence of an utterance helped the language identification performance even with the presence of phonetic recognition errors.

Despite the fact that the language model was the most dominate component of the ALI system, this thesis showed that additional information, such as prosodic and acoustic information, can also be useful for language identification. When the acoustic and prosodic models were incorporated into the ALI system to support the language model, the system's accuracy increased from 41.7% to 48.6%. Additionally, despite the simplicity of the modeling of the prosodic features, the prosodic model proved to be more useful for language identification than the acoustic model.

5.2 Assessment of System Performance

As discussed in Chapter 1, it is very difficult to compare an ALI system to the *state of the art* in ALI because there have been very few studies which utilize a comparable evaluation task. With the recent release of the OGI Multi-Language Telephone Speech Corpus into the public domain, there now exists a common data set from which meaningful comparisons of different ALI approaches can be made. NIST is currently coordinating a series of ALI evaluations utilizing the OGI corpus to compare the approaches of eleven different research efforts.[1] To date, two studies have published preliminary results using the OGI corpus. These studies were conducted by Muthusamy and Cole [28] and by Zissman [36]. Table 5.1 shows how the results reported in this thesis compare to the results of their systems.[2] As can be seen, the system developed in this thesis is competitive with the other two systems.[3]

Authors of Study	Date	System Accuracy
Hazen	August, 1993	48.6%
Muthusamy and Cole	September, 1992	47.7%
Zissman	April, 1993	46.0%

Table 5.1: Summary of results using the OGI Multi-Language Telephone Speech Corpus

The performance of the system is quite promising considering the difficulty of the task. The OGI corpus contains many features which can adversely affect the system's performance. Some of the difficulties of the corpus include:

- The data set is currently unlabeled making fully supervised training impossible.
- The data set was collected over many different channels of varying qualities.
- The data set was sampled at a rate of only 8 kHz, limiting its bandwidth.
- A large portion of the data contains completely unconstrained speech.

[1] NIST is the National Institute of Standards and Technology.

[2] Both groups are currently continuing their ALI research and improved results can be expected.

[3] The training and test sets extracted from the OGI corpus were identical for all three systems.

5.3 Future Work

5.3.1 System Improvements

Though the system developed in this thesis has proven to be competitive with other current ALI systems, there are still many improvements that can be made. In particular, future research will attempt to satisfy the following goals:

- Improve the phonetic recognition component of the system.
- Investigate methods for channel normalization.
- Discover more useful segment-based features for acoustic modeling.
- Develop modeling schemes to capture the correlation between the F0 contour and the segment durations.
- Develop modeling schemes to better capture the dynamic characteristics of the F0 contour.
- Examine different approaches for system integration.

As shown in Chapter 3, the performance of the language modeling component of the system is very dependent on the quality of the representation of the underlying phonetic sequence. Therefore, it is important for the phonetic recognizer used by the ALI system to be as accurate as possible. Since neither of the phonetic recognizers used in this thesis was trained in a fully supervised fashion, large improvements in the phonetic recognition accuracy may not be possible until fully supervised training can be implemented. This may be feasible in the near future when the phonetic transcriptions of the OGI data become available.

Because the OGI corpus was collected over the telephone lines using a different channel for every speaker, the acoustic qualities of the speech can vary significantly from speaker to speaker. Therefore, the ALI system should account for the acoustic differences between the channels in its modeling schemes to help avoid any channel dependencies which may arise. The ALI design in this thesis does not account for the channel differences. Therefore, future work will investigate methods, such as blind deconvolution, for channel normalization.

The acoustic model used in this thesis attempts to model the acoustic information of the different phonemes in each language using segment-based feature vectors. The feature vectors that were used were relatively simple in nature; they contained the values of the MFCCs and delta MFCCs averaged over the length of a segment. The acoustic model may be improved by using a different set of features. In fact, the acoustic features that are useful for language identification may be quite different from the features that are useful for phonetic recognition. It has been shown that useful segment-based acoustic measurements for phonetic recognition can be discovered in an automatic fashion [30]. It may be possible to automatically discover useful segment-based measurements for language identification in a similar fashion. Thus, future work will include attempting to discover more useful segment-based acoustic features.

The prosodic model used in this thesis attempts to capture the F0 and segment duration information using simple statistical properties. The independence assumptions that were made may in fact be hurting the performance of the prosodic model. The first major assumption was to treat the segment durations and the F0 contour as independent entities. Because of the correlations that may exist between the segment durations and the F0 contour in the creation of the stress or tone of a segment, this assumption may be inappropriate. The second major assumption was to treat each frame of the F0 contour as independent. This assumption eliminates almost all of the dynamic information contained in the F0 contour. Thus, future work will include efforts to create models which can account for the correlations between the F0 contour and segment durations as well as the dynamic nature of the F0 contour over time.

An additional assumption that was used in the prosodic model was that the unvoiced frames of the utterance carried no useful information and could be ignored. Although a preliminary experiment which incorporated the probability of voicing parameter into the F0 model did not yield any improved performance in the F0 model, this is an assumption which also requires further study.

As mentioned in Chapter 4, new procedures for integrating the different models into the complete system should be investigated. The static scaling method described in Chapter 3 does not account for the possibility that some models may contribute significantly more useful information than others as the length of an utterance is increased. Thus, methods for dynamically changing the scaling factors for each model as the length of an utterance increases should be investigated.

The final system also did not use the same linguistic sequence $\hat{C}$ for each of the models. The language and acoustic models used a $\hat{C}$ which was represented with 59 phonetic classes while the segment duration model used a $\hat{C}$ which was represented with 29 phonetic classes. The fact that the models are not modeling the same probability space may be hurting the system. Thus, better methods need to be developed to find the single phonetic representation of $\hat{C}$ which optimizes the system's performance.

5.3.2 Incorporation into a Multi-Lingual System

As mentioned in Chapter 1, an ALI system can be utilized as a component within a larger multi-lingual system. As a testbed for multi-lingual research, a multi-lingual information retrieval system is currently under development in the Spoken Language

Systems group at MIT. This system, known as the multi-lingual VOYAGER system, is designed to provide travel information for the city of Cambridge [38, 39, 42]. VOYAGER currently has the capability to understand queries in either English or Japanese [7], and is being ported to French, Italian and German.

Within the multi-lingual VOYAGER domain, ALI can be performed as a two step process. The first step is to perform a *fast match* to provide an ordered list of possible language candidates. The second step is to utilize the speech recognizer of the top choice language candidate to attempt to decipher the utterance. If the speech recognizer for the top-choice language fails to understand the utterance, the utterance is passed to the recognizer for the second choice, and so forth, until the system is able to understand the input query. In this scenario, the ALI design described in this thesis could be used to provide the language identification fast match.

When the ALI system is incorporated into a system such as multi-lingual VOYAGER, the tradeoff between accuracy and efficiency is very important. Since higher level knowledge of each language is available, the entire system should be able to perform nearly flawless language identification for sentences within it domain. The goal of the system is thus shifted from accuracy to speed. In this two-tiered approach to language identification, the optimal solution may involve creating an ALI fast match component which sacrifices accuracy for the sake of efficiency.

Thus, future work will also be directed at incorporating the ALI design described in this thesis into the multi-lingual VOYAGER system. This will involve a careful study of the tradeoff between the design's computational efficiency and its language identification accuracy.

Appendix A

Families of OGI Languages

Figure A.1 shows a tree describing the ten languages in the OGI corpus in terms of their linguistic origins [31]. It should be noted that the structure of the tree in Figure A.1 is derived from only one of many different hypotheses that linguists have proposed to describe the development of the different languages of the world. Furthermore, to date linguists have been unable to determine whether or not any of the approximately 30 primary language families of the world were developed from a single common source or whether these language families came into existence independently. Thus, it is only for aesthetic reasons that the structure of the tree in Figure A.1 is shown with the five primary language families originating from a common node.

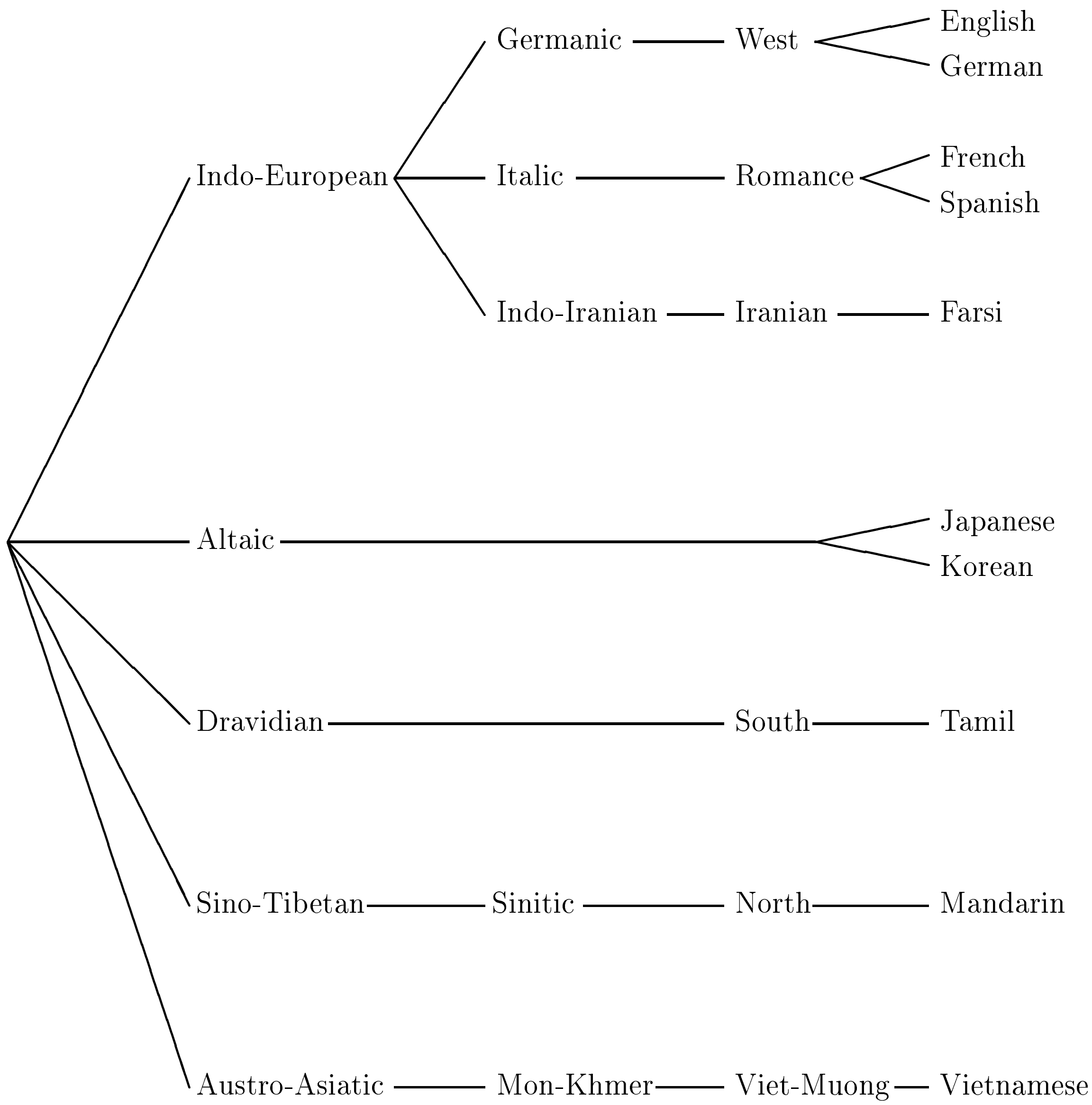

Figure A.1: Language family tree of the 10 languages in the OGI corpus

Appendix B

Phone Sets of OGI Languages

Table B.1 displays the phones which are used in each of the ten languages used in this thesis. The phones are written using the standard International Phonetic Association (IPA) alphabet. The table is created from the language specific phonetic lists compiled by Ruhlen [31]. These lists include all of the primary realizations of the phonemes of the language. However, as Ruhlen states, the lists do not always contain context specific allophones. For example, Ruhlen does not list the flap [ɾ] as a phone in American English because it is simply a context dependent allophone for the phoneme /t/. Ruhlen also does not include diphthongs in his lists (although they are included for American English). Despite the incompleteness of the lists, they still provide a general idea of which sounds can be expected in each of the ten languages in the OGI database.

Phonetic class	Phones in each language									
	Eng	Far	Fre	Ger	Jap	Kor	Man	Spa	Tam	Vie
Vowels	iʲ ɪ eʲ ə æ ɛ a ʌ aʲ aʷ ɔ ɔʲ oʷ ʊ u	i e æ a o u	i y e ø ɛ œ a ɔ o u ə æ̃ ɑ̃ õ	i y e ø ɑ ə o u	i e ɑ o u	i e ɛ ɒ o ɤ u ɯ	i y e a o u	i e a o u	i e a o u ẽ ã õ ũ	i e ɛ æ ɐ ɔ ʌ o ɤ ɯ u
Stops	pʰ b tʰ d kʰ g	pʰ b t̠ʰ d̠ kʰ g ɢ ʔ	p b t̠ d̠ k g	pʰ b t̠ʰ d̠ kʰ g	p b t̠ d̠	p pʰ pˀ t̠ t̠ʰ t̠ˀ k kʰ kˀ	p pʰ t̠ t̠ʰ k kʰ	p b t̠ d̠ k g	p b t̠ d̠ ṭ ḍ t k g	p ɓ t̠ʰ t ɗ c k ʔ
Affricates	ʧ ʤ	ʧ ʤ		pᶠ t̠ˢ	ʧ ʤ t̠ˢ	ʧ ʧʰ ʧˀ	t̠ˢ t̠ˢʰ ṭˢ ṭˢʰ cᶜ̧ cᶜ̧ʰ	ʧ	ʧ ʤ	
Fricatives	f v θ ð s z ʃ ʒ h	f v s z ʃ ʒ χ h	f v s z ʃ ʒ	f v s z ʃ ʒ ç x h	s z ʃ ʒ h	s sˀ h	f s̠ ṣ ç χ	f θ s x	f s ç	f s̠ z̠ x ɣ h
Nasals	m n ŋ	m n̠	m n̠ ɲ	m n̠	m n̠	m n̠ ŋ	m n̠ ŋ	m n̠ ɲ	m n̠ n ṇ ŋ	m n̠ ɲ ŋ
Liquids	l ɹ̣	l r	l ʀ	l̠ ʀ	l ɾ	l	l ɹ̣	l̠ ʎ r ɾ	l ḷ r	l̠
Glides	j w	j w	j ɥ w	j w	j w	j w	j ɥ w	j w	j w	j w ɯ̯ ɒ̯

Table B.1: Phone Sets of Languages in OGI Database

Bibliography

[1] Deidre Cimarusti and Russell B. Ives. Development of an automatic identification system of spoken languages: Phase I. In *Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing*, pages 1661–1663. IEEE, 1982.

[2] Nancy A. Daly and Victor W. Zue. Acoustic, perceptual, and linguistic analyses of intonation contours in human/machine dialogues. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, November 1990.

[3] Nancy A. Daly and Victor W. Zue. Statistical and linguistic analyses of F_0 in read and spontaneous speech. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, October 1992.

[4] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[5] L. Fissore, P. Laface, and G. Micca. Comparison of discrete and continuous HMMs in a CSR task over the telephone. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 253–256. IEEE, 1991.

[6] Jerry T. Foil. Language identification using noisy speech. In *Proceedings of the 1986 International Conference on Acoustics, Speech, and Signal Processing*, pages 861–864. IEEE, 1986.

[7] James Glass, David Goodine, Michael Phillips, Shinsuke Sakai, Stephanie Seneff, and Victor Zue. A bilingual VOYAGER system. In *Proceedings of the 1993 European Conference on Speech Communication and Technology*, September 1993.

[8] James R. Glass. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*. PhD thesis, Massachusetts Institute of Technology, May 1988.

[9] James R. Glass and Victor W. Zue. Multi-level acoustic segmentation of continuous speech. In *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing*, pages 429–432. IEEE, 1988.

[10] Fred J. Goodman, Alvin F. Martin, and Robert E. Wohlford. Improved automatic language identification in noisy speech. In *Proceedings of the 1989 International Conference on Acoustics, Speech, and Signal Processing*, pages 528–531. IEEE, 1989.

[11] Arthur S. House and Edward P. Neuburg. Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustic Society of America*, 62(3):708–713, September 1977.

[12] Ken-ichi Iso and Takao Watanabe. Large vocabulary speech recognition using neural prediction model. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 57–60. IEEE, 1991.

[13] R. B. Ives. A minimal rule AI expert system for real-time classification of natural spoken languages. In *Proceedings of the Second Annual Artificial Intelligence and Advanced Computer Technology Conference*, pages 337–340, 1986.

[14] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing*, pages 109–112, April 1990.

[15] F. Jelinek. Self-organized language modeling for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, chapter 8, pages 450–506. Morgan Kaufmann Publishers, 1990.

[16] Lori F. Lamel and Jean-Luc Gauvain. Cross-lingual experiments with phone recognition. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1993.

[17] Lori F. Lamel, Robert H. Kassel, and Stephanie Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 100–109, 1986.

[18] R. G. Leonard. Language recognition test and evaluation. Technical Report RADC-TR-80-83, Air Force Rome Air Development Center, March 1980.

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

[20] R. G. Leonard and G. R. Doddington. Automatic language identification. Technical Report RADC-TR-75-264, Air Force Rome Air Development Center, October 1975.

[21] R. G. Leonard and G. R. Doddington. Automatic language discrimination. Technical Report RADC-TR-78-5, Air Force Rome Air Development Center, January 1978.

[22] Hong C. Leung, I. Lee Hetherington, and Victor W. Zue. Speech recognition using stochastic explicit-segment modeling. In *Proceedings of the Second European Conference on Speech Communication*, 1991.

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

[24] John Makhoul, Salim Roucus, and Herbert Gish. Vector quantization in speech coding. *Proceedings of the IEEE*, 73(11):1551–1588, November 1985.

[25] Helen Mei-Ling Meng. The use of distinctive features for automatic speech recognition. Master's thesis, Massachusetts Institute of Technology, May 1991.

[26] P. Mermelstein and S. Davis. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), August 1980.

[27] Yeshwant K. Muthusamy, Ronald A. Cole, and Murali Gopalakrishnan. A segment-based approach to automatic language identification. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 353–356. IEEE, 1991.

[28] Yeshwant K. Muthusamy, Ronald A. Cole, and Beatrice T. Oshika. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, 1992.

[29] Yeshwant K. Muthusamy, Ronald A. Cole, and Beatrice T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 1007–1010, 1992.

[30] Michael Phillips and Victor Zue. Automatic discovery of acoustic measurements for phonetic classification. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, October 1992.

[31] Merritt Ruhlen. *A Guide to the Languages of the World*. Stanford University, 1976.

[32] Michael Savic, Elena Acosta, and Sunil K. Gupta. An automatic language identification system. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 817–820. IEEE, 1991.

[33] B. G. Secrest and G. R. Doddington. An integrated pitch tracking algorithm for speech systems. In *Proceedings of the 1983 International Conference on Acoustics, Speech, and Signal Processing*, pages 1352–1355. IEEE, 1983.

[34] Masahide Sugiyama. Automatic language recognition using acoustic features. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 813–816. IEEE, 1991.

[35] Jacqueline Vaissière. Language-independent prosodic features. In Anne Cutler and D. Robert Ladd, editors, *Prosody: Models and Measurements*, chapter 5, pages 53–66. Springer-Verlag, 1983.

[36] Marc A. Zissman. Automatic language identification using Gaussian mixture and hidden Markov models. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, pages 399–402. IEEE, 1993.

[37] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. Recent progress on the SUMMIT system. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, June 1990.

[38] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. The VOYAGER speech understanding system: A progress report. In *Proceedings of the Second DARPA Speech and Natural Language Workshop*, October 1989.

[39] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. The VOYAGER speech understanding system: Preliminary development and evaluation. In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing*, April 1990.

[40] Victor Zue, James Glass, David Goodine, Michael Phillips, and Stephanie Seneff. The SUMMIT speech recognition system: Phonological modeling and lexical access. In *Proceedings of the 1990 International Conference on Acoustics, Speech and Signal Processing*, April 1990.

[41] Victor Zue, James Glass, Michael Phillips, and Stephanie Seneff. The MIT SUMMIT speech recognition system: A progress report. In *Proceedings of the DARPA Speech and Natural Language Workshop*, February 1989.

[42] Victor W. Zue, James R. Glass, Dave Goddeau, David Goodine, Hong C. Leung, Michael K. McCandless, Michael S. Phillips, Joseph Polifroni, Stephanie Seneff, and Dave Whitney. Recent progress on the MIT VOYAGER spoken language system. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, November 1990.

[43] Victor. W. Zue and Stephanie Seneff. Transcription and alignment of the TIMIT Database. In *Proceedings of the Second Meeting on Advanced Man-Machine Interface through Spoken Language*, 1988.

[44] Victor W. Zue, Stephanie Seneff, and James R. Glass. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9(4):351–356, August 1990.