Empty Page

Segment-based automatic language identification

Timothy J. Hazen and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

Received

Footline title: Automatic language identification

Abstract

This paper discusses the formulation, development and analysis of a segment-based approach to the Automatic Language Identification (LID) problem. This system utilizes phonotactic, acoustic-phonetic and prosodic information within a unified probabilistic framework. The implementation of this framework allows the relative contributions of different sources of information to be determined empirically, as well as providing the mechanism for combining them within one system. The system has been evaluated using the OGI Multi-Language Telephone Speech Corpus and the results are competetive with other current LID systems. The results have also indicated that, while the phontotactic information of a spoken utterace is the most useful information for LID, acoustic-phonetic and prosodic information can be useful for increasing a system's accuracy, especially when the utterance is short.

PACS number: 43.72

1 Introduction

Automatic language identification (or LID) refers to the task of identifying the language being spoken by a person. For more than twenty years, interests and needs in the intelligence community have provided a major impetus for research in LID (Leonard and Doddington, 1974, 1975, 1978; Leonard, 1980). More recently, LID research has been enjoying a renaissance, spurred by research activities in multi-lingual speech recognition and understanding for which an efficient means for identifying the language being spoken has definite benefits. For example, telephone companies can provide better services to customers speaking different languages if an LID front-end can route calls to the appropriate operators. Similarly, a multi-lingual spoken language system (Glass *et al.*, 1993; Flammia *et al.*, 1994) can understand and respond in the user's native language if it can first identify the language being spoken. Machine translation systems can utilize LID to relieve the burden of requiring users to specify the language they are speaking before each turn. For a thorough summary of the research that has been performed in automatic language identification over the last 20 years, please refer to (Zissman, 1996).

Because languages of the world can be distinguished amongst one another by their own unique vocabulary and linguistic constructs, a spoken language can be identified, in principle, by passing the speech signal through a set of speech recognition/understanding systems running in parallel, each capable of deciphering a particular language of interest. Language identification can then be achieved implicitly by choosing the language of the system that yields the best score. Viewed in this manner, a reliable speech recognition/understanding system for all spoken languages should provide near perfect language identification when these systems are used jointly.

However, there are several reasons why such an approach may be impractical. Building a multi-lingual recognition/understanding system as described above requires extensive knowledge about the acoustic-phonetic, lexical, and linguistic rules for each of the languages of interest. Even for restricted domains, this knowledge may require a tremendous effort to acquire. The resulting systems may not be easily portable, since they will only perform well when the vocabulary and linguistic rules of the languages in question are well-specified. As such, porting such systems to new languages would be laborious and time consuming. Furthermore, it may be too computationally expensive to incorporate all of this knowledge even if it were available. Thus, the goal of LID research to date has generally been to develop dependable language identification methods which do not rely upon higher level knowledge of the languages involved, but rather utilize only the information that is available directly from the waveform.

It has been observed that humans often can identify the language of an utterance even when they have no working linguistic knowledge of that language (Muthusamy and Cole, 1992), suggesting that they are able to learn and recognize language-specific patterns directly from the signal. In the absence of higher level knowledge of a language, a listener presumably relies on lower level constraints such as acoustic-phonetics (i.e., the inventory and characteristics of sound units), phonotactics (i.e., the sequential constraints on sound patterns), and prosody (i.e., the supra-segmental properties).

The constraining power of these low-level characteristics for LID was first demonstrated in a feasibility study by House and Neuburg (House and Neuburg, 1977). Their results offered the hope that very simple phonetic language models can be powerful tools for language identification. In this paper, we describe a LID system that is primarily built upon House and Neuburg's initial ideas. The speech signal is first segmented and classified into sequences of phonetic classes and the phonotactic properties of the resulting sequences are modeled statistically. Since other information besides phonotactics may also be useful for language identification, as has been demonstrated by other investigators (Muthusamy *et al.*, 1991, Muthusamy and Cole, 1992; Itahashi *et al.*, 1994) our system will supplement the phonotactic information with prosodic and acoustic-phonetic information, all within one unified framework.

In this paper, we will first derive the theoretical framework for our LID system. This will be followed by a description of the system architecture and the resulting implementation. We contrast this system's design with the current state of the art systems. Performance of our system on the OGI Multi-Language Telephone Speech Corpus (Muthusamy *et al.*, 1992; Cole *et al.*, 1994) under varying parameter settings will be presented and analyzed.

2 Theory

2.1 Overview

An understanding of the characteristics of spoken language which are most useful for discriminating among languages is essential to the development of an LID system. An LID system must 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.

The language-discriminating information contained in the signal can be segmental and prosodic (e.g., suprasegmental). The segmental information can be acoustic-phonetic, which relates to the manner in which phones are realized acoustically by a speaker, or phonotactic, which relates to the higher-level rules governing the sequences of phones which are allowed within a language. Prosodic information is encoded in the fundamental frequency, intensity and duration variations that span across segments. While the segmental and prosodic information may also carry higher-level linguistic constraints, our working assumption is that knowledge of this higher-level information will not be needed to identify the language of the utterance. In this section, we will briefly outline the theoretical framework for our LID system. Interested readers are referred to (Hazen, 1993) for a more detailed treatment.

2.2 Probabilistic Framework

The system described in this paper utilizes a segment-based probabilistic framework. 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 LID 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 representing the acoustic information at a fixed frame rate. For this derivation, we will assume that two specific types of information are extracted from the waveform for each time frame: 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 when 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 $\vec{\mathbf{a}} = \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$ be the sequence of m vectors which represent the wide-band spectral information of a spoken utterance and let $\vec{\mathbf{f}} = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_m\}$ be the sequence of m vectors which represent the voicing information of a spoken utterance. Throughout this paper, the wide-band spectral information contained in $\vec{\mathbf{a}}$ will be referred to as the acoustic information and the information in $\vec{\mathbf{f}}$ will be referred to as the F0 information.

Next, assume that we can utilize a phonetic speech recognizer to extract the most likely sequence of phonetic elements contained in the utterance. Let the recognized phonetic sequence, containing p phonetic elements, be represented as $C = \{c_1, c_2, \ldots, c_p\}$ where each c is represented with a specific phonetic element. For a segment-based approach, as is being pursued in our group, the concept of segmentation of the input speech must be incorporated into the probabilistic framework. Thus, let $S = \{s_1, s_2, \ldots, s_{p+1}\}$ represent the segmentation for the phonetic string C where each s represents the location of a segment boundary.

Given, the acoustic information $\vec{\mathbf{a}}$, the fundamental frequency $\vec{\mathbf{f}}$, the most likely phonetic sequence C and the segmentation S the most likely language is found using the following expression:

$$\arg\max_i \Pr(L_i \mid C, S, \vec{\mathbf{a}}, \vec{\mathbf{f}}). \tag{1}$$

Using standard probability theory, this expression can be equivalently written as:

$$\arg\max_i \Pr(\vec{\mathbf{a}} \mid C, S, \vec{\mathbf{f}}, L_i) \Pr(S, \vec{\mathbf{f}} \mid C, L_i) \Pr(C \mid L_i) \Pr(L_i). \tag{2}$$

The four probability expressions in (2) are organized in such a way that prosodic and phonetic information are contained in separate terms. In modeling, these terms become known as:

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

The phonetic information is contained in two separate models: the phonetic acoustic model and the phonetic language model. In subsequent sections these models will

simply be referred to as the acoustic model and the language model. The acoustic model accounts for the different acoustic realizations of the phonetic elements that may occur across languages, whereas the language model accounts for the probability distributions of the phonetic elements and the phonotactic constraints within each language. 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. This organization provides a useful structure for evaluating the relative contributions towards language identification that phonotactic, acoustic-phonetic, and prosodic information provide.

3 System Design and Implementation

3.1 General System Architecture

The architecture of our LID system implementing the segment-based probabilistic framework described in Section 2 is shown in Figure 1. It is realized as a series of four components: a preprocessor, a phonetic recognizer, a fast match language identifier, and a language verifier. The preprocessor receives the raw acoustic waveform as its input and transforms it into the frame-based feature vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{f}}$. The phonetic recognizer receives the acoustic information in $\vec{\mathbf{a}}$ as its input and finds the best phonetic hypothesis and segmentation, C and S. The language identifier uses $\vec{\mathbf{a}}$, $\vec{\mathbf{f}}$, C and S to construct a candidate list of the most likely languages. The language verifier then utilizes language dependent speech recognizers to verify the languages provided by the language identifier. In this section we summarize the design choices made for each of the language identification components. Interested readers are referred to (Hazen, 1993) and (Hazen and Zue, 1994) for a more detailed description.

Our approach differs from similar approaches taken by Zissman (Zissman and Singer, 1994; Zissman, 1995) and by Kadambe and Hieronymus (Kadambe and Hieronymus, 1994, 1995) in that our system uses one language-*independent* phonetic recognizer instead of a set of language-*dependent* phonetic recognizers. Zissman has shown empirically that increasing the number of language-dependent recognizers increases performance in his system (Zissman, 1995). However, the increased performance comes at the expense of increased computation. It is for computational efficiency that we have adopted this simpler architecture.

3.2 Preprocessing

The acoustic vector $\vec{\mathbf{a}}$ is represented with mel-frequency cepstral coefficients (MFCC's) (Mermelstein and Davis, 1980). A set of 14 MFCC's (including the energy term) are computed for each utterance every 5 ms using a 25.6 ms Hamming window and a 256-point discrete Fourier transform. To compensate for the varying acoustic properties of the different channels encountered in the OGI data, a blind deconvolution channel normalization scheme is also employed. For each utterance, the average value of each of the fourteen MFCC's is computed over the length of the utterance. This average is then subtracted from the MFCC value of each frame in the utterance. In addition to the MFCC's, 14 delta MFCC's are also computed.

The voicing information contained in the vector $\vec{\mathbf{f}}$ is extracted from the acoustic

signal using a pitch detector originally devised by Secrest and Doddington (Secrest and Doddington, 1983) and incorporated as part of the FORMANT program in Entropic's ESPS package. 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, again 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.

3.3 Phonetic Recognition

The phonetic recognition component decodes the acoustic information into a string of phonetic events. In our system, phonetic recognition is performed by SUMMIT, a segment-based speech recognition system developed in our group (Zue *et al.*, 1989, 1990). SUMMIT utilizes a hierarchical segmentation algorithm to provide the segmentation search space. An ordered list of potential phone candidates and their respective likelihoods are produced for each potential segment. The phone likelihoods are obtained from mixture Gaussian density functions for each phone which model segment-based feature vectors. A search algorithm is applied to the segmentation and phone search space to find the most likely strings of phones. In our implementation, we used 87 language-independent phonetic units. Included in this set are several different silence and noise units. The 87 phones were the result of hand clustering over 900 unique labels which exist in the transcriptions of the training data. Similar phones, such as [iː] and [i], were collapsed into a single class to ensure an adequate amount of training data for each class while still maintaining a richness in phonetic description. The complete set of phonetic units used by the system is shown in Table 1. To prevent the recognizer from being biased towards the phones of particular languages, no phonetic language model was used by SUMMIT (i.e., the *a priori* probabilities of the phones were presumed to be uniform).

One primary difference between the training of the standard SUMMIT system and the system used in these experiments lies in the manner in which the segment boundary scoring parameters were selected. During standard training of the SUMMIT phonetic recognizer, the segment boundary scoring parameters are chosen to optimize the phonetic recognition performance on development data. For our experiments, these parameters were chosen to optimize the language identification accuracy of the language model which utilizes the output of the phonetic recognizer. This optimization process favored the removal of deletion errors at the expense of increased insertion

errors, suggesting that segment deletion, which removes information, can be more harmful than the insertion of either extraneous information or additional noise.

3.4 Language Identification

3.4.1 General Framework

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

$$\arg\max_i \Pr(\vec{\mathbf{a}} \mid C, S, \vec{\mathbf{f}}, L_i) \Pr(S, \vec{\mathbf{f}} \mid C, L_i) \Pr(C \mid L_i) \Pr(L_i). \tag{3}$$

For our experiments, the *a priori* language probability distribution was assumed to be uniform and hence ignored. As is standard, the expression was realized using its logarithmic form as follows:

$$\arg\max_i \left(\log \Pr(\vec{\mathbf{a}} \mid C, S, \vec{\mathbf{f}}, L_i) + \log \Pr(S, \vec{\mathbf{f}} \mid C, L_i) + \log \Pr(C \mid L_i) \right). \tag{4}$$

Figure 2 illustrates the components which are utilized in the realization of the expression in (4). Essentially, each language has its own language, acoustic, and prosodic models. The incoming utterance is scored by the models of each language and the candidate languages are then ranked by their respective scores. Figure 2 also shows one additional component not represented in (4). This component is the set of model weights w_1, w_2, and w_3. Ideally, these weights should all simply be set to one. However, as will be discussed later, these weights are necessary to avoid having one model dominate the total score for each language.

3.4.2 System Training

Each term in (4) must be modeled using the transcription of the utterance as represented by the phonetic string C and the segmentation S. Standard speech recognition techniques which require a transcription for supervised training, such as n-gram and mixture Gaussians, are used to model the terms. To provide this transcription each training utterance is passed through the phonetic recognizer to produce an automatically generated transcription. The primary deviation from the ways these models are usually trained stems from the fact that, in our case, C and S do not represent the actual transcription of the underlying string of phones, but rather the output of a recognizer which is prone to errors. Thus the language, acoustic, and prosodic models

which are eventually generated do not model the actual characteristics of each language, but rather the characteristics of each language after its utterances have been corrupted by the noise from the imperfect recognizer. For accurate language identification to occur, it is necessary for the noise generated by the phonetic recognizer to be accounted for in the modeling process since it will exist in the test data.

3.4.3 Language Model

The language model refers to the expression $\Pr(C \mid L_i)$. The language model is potentially the most important element of the system. House and Neuburg showed that simple n-gram language models applied to error free sequences of phonetic elements as general as broad phonetic classes can reliably identify the language of an utterance (House and Neuburg, 1977). However, the language identification capabilities of an n-gram are degraded when the actual string of phonetic events is corrupted with errors.

The language model used for our experiments was an interpolated trigram model (Jelinek, 1990). The interpolated trigram can be expressed as

$$\hat{\Pr}(c_i \mid c_{i-1}, c_{i-2}) = \lambda_2 \Pr(c_i \mid c_{i-1}, c_{i-2}) + (1 - \lambda_2)\left(\lambda_1 \Pr(c_i \mid c_{i-1}) - (1 - \lambda_1)\Pr(c_i)\right) \tag{5}$$

where λ_1 and λ_2 are weights which depend on the phones preceding c_i. Specifically, λ_1 is expressed as

$$\lambda_1 = \frac{k_{c_{i-1}}}{k_{c_{i-1}} + K_1} \tag{6}$$

where $k_{c_{i-1}}$ is the number of exemplars of c_{i-1} in the training set and K_1 is a constant. Similarly, λ_2 is expressed as

$$\lambda_2 = \frac{k_{c_{i-1},c_{i-2}}}{k_{c_{i-1},c_{i-2}} + K_2}. \tag{7}$$

The constants K_1 and K_2 were chosen empirically to optimize the language identification performance of the interpolated trigram model on development test data extracted from the training set. The optimization process set the values of K_1 and K_2 to 350 and 800 respectively.

3.4.4 Acoustic Model

The expression $\Pr(\vec{\mathbf{a}} \mid \vec{\mathbf{f}}, S, C, L_i)$ is called the acoustic model, which is used to capture information about the acoustic realizations of each of the phones used in each

language. To simplify the modeling, the acoustic information $\vec{\mathbf{a}}$ is assumed to be independent of the fundamental frequency information $\vec{\mathbf{f}}$, and each segment is considered to be independent of all other segments. Like others before us, we make the statistical independence assumption realizing that it is, in all likelihood, difficult to justify. However, it is our belief that that such an assumption will be less troublesome for us because we are dealing with segments rather than frames. With these assumptions, the acoustic model can be expressed as

$$\Pr(\vec{\mathbf{a}} \mid \vec{\mathbf{f}}, S, C, L_i) = \Pr(\vec{\mathbf{a}} \mid S, C, L_i) = \prod_{k=1}^{m} \Pr(\vec{a}_k \mid c_k, L_i) \tag{8}$$

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. In our case, each $\vec{a}_k$ contains 14 MFCC's and 14 delta MFCC's as averaged over the length of the segment.

Using the above assumptions, continuous probability density functions which model the segment-based acoustic feature vectors for each phone in each language can be used for the acoustic model. The acoustic feature vectors are modeled with mixtures of diagonal Gaussian density functions. To create the mixture Gaussian model for each phone in each language, the set of Gaussian 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 ensure proper amounts of training data for each mixture of Gaussians, the number of Gaussians used to model each phonetic class is determined from the equation

$$n_g = \begin{cases} n_{max} & \text{if } k/100 > n_{max} \\ \lceil k/100 \rceil & \text{otherwise} \end{cases} \tag{9}$$

where n_g 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. The value of n_{max} was chosen to be 16 based upon prior experiments (Hazen, 1993).

3.4.5 Prosodic Model

The prosodic model refers to the expression $\Pr(S, \vec{\mathbf{f}} \mid C, L_i)$. Ideally, this model can be used to capture the differences among languages that exist in the prosodic structure of utterances. While useful and reliable methodologies are available for modeling acoustic and phonetic information, well-developed techniques for automatically capturing

and understanding word- and sentence-level prosodic information remains elusive. Therefore, our prosodic model only captures simple statistical information about the fundamental frequency and segment duration information of an utterance.

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

$$\Pr(S, \vec{\mathbf{f}} \mid C, L_i) = \Pr(\vec{\mathbf{f}} \mid S, C, L_i) \Pr(S \mid C, L_i). \tag{10}$$

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. This simplification and the independence assumptions which will further be made clearly ignore much of the information that should be captured by the prosodic model. An early experiment combining the fundamental frequency and segment duration models into a single segment-based prosodic model did not produce satisfactory results (Hazen and Zue, 1994). One potential way to improve these models is to consider more descriptive measurements such as those proposed in (Muthusamy and Cole, 1992) and (Itahashi *et al.*, 1994).

Fundamental Frequency Model

The expression $\Pr(\vec{\mathbf{f}} \mid S, C, L_i)$ captures the information available in the F0 contour of an utterance. While there may be correlation between the F0 contour and the durations of the segments in the utterance, this correlation is ignored in order to simplify the modeling of the F0 contour. Thus, $\vec{\mathbf{f}}$ will be considered independent of S and C. With these assumptions the fundamental frequency model can be simplified as follows:

$$\Pr(\vec{\mathbf{f}} \mid S, C, L_i) = \Pr(\vec{\mathbf{f}} \mid L_i). \tag{11}$$

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 is considered to be statistically independent. With this assumption the F0 model can be written as

$$\Pr(\vec{\mathbf{f}} \mid L_i) = \prod_{k=1}^{m} \Pr(\vec{f}_k \mid L_i) \tag{12}$$

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 (12) only includes the frames which are voiced. The expression in (12) can be modeled with a mixture of full covariance Gaussian probability density functions. Based on prior experiments (Hazen, 1993), the number of full covariance density functions utilized in the mixture Gaussian model for each language was chosen to be 9.

Segment Duration Model

The expression $\Pr(S \mid C, L_i)$ captures the segment duration information in a utterance. While there may be very useful information contained in 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 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

$$\Pr(S \mid C, L_i) = \prod_{k=1}^{m} \Pr(d_k | c_k, L_i) \tag{13}$$

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

The expression $\Pr(d_k|c_k, L_i)$ can be modeled with a mixture of Gaussian models. As with the acoustic model, the number of Gaussians used to model each phone in each language is determined by the equation

$$n_g = \begin{cases} n_{max} & \text{if } k/30 > n_{max} \\ \lceil k/30 \rceil & \text{otherwise} \end{cases} \tag{14}$$

where n_g is the number of Gaussians used in the mixture Gaussian model of a particular phone for a particular language, n_{max} is the maximum number of Gaussians allowed in each mixture, and k is the number of training vectors available for the phone in that particular language. Based on prior experiments (Hazen, 1993), the value of n_{max} was chosen to be 4.

3.4.6 System Integration

Finally, each of the individual models must be integrated into the complete LID system. i.e., the likelihood scores from each individual model for an utterance must be combined to provide one likelihood score for each language. Using the probabilistic

framework, this can is accomplished with the following expression:

$$\arg\max_i \left(\log \Pr(\vec{\mathbf{a}} \mid C, S, \vec{\mathbf{f}}, L_i) + \log \Pr(S, \vec{\mathbf{f}} \mid C, L_i) + \log \Pr(C \mid L_i) \right). \qquad (15)$$

However, as documented in (Hazen, 1993), 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 acoustic model score. To compensate for this effect, the score from each model is multiplied by a weighting factor, as shown in Figure 2. A hill-climbing optimization procedure is utilized to find an adequate set of weighting factors. This procedure adjusts the weights for the various models to optimize the language identification performance of the system on development data jackknifed from the training set. Because the weighting factors only need to provide a means of adjusting the *relative* scores of each model and not the absolute scores, the weight of the language model was pinned to a value of one while the weights for the acoustic, duration, and F0 models were allowed to vary during the iterative optimization process. The weighting factors were also optimized for different test utterance lengths. Figure 3 shows the weighting factors found by the hill-climbing procedures for the acoustic, duration and F0 models. Note that as the test utterance length increases, the weights of the acoustic, duration and prosodic models generally decrease. This effectively gives the language model more weight for longer utterances. Additionally it should be noted that, while the duration and prosodic model contribute to the total score even as the utterance length increases, the acoustic model effectively contributes nothing to the total score for longer utterances.

4 Experimental Results and Discussion

The effectiveness of our LID system is empirically determined using the OGI Multi-Language Telephone Speech Corpus (Muthusamy *et al.*, 1992b; Cole *et al.*, 1994). This corpus contains utterances collected over the telephone lines from native speakers of 11 different languages. For our experiments, we used a training set containing 5,987 topic-specific as well as unconstrained utterances. Of these, 471 utterances were accompanied by time-aligned phonetic transcriptions. The primary test set for our experiments contained 187 utterances as selected by The National Institute of Standards and Technology (NIST) for their March 1994 LID evaluation. These utterances were all a minimum of 30 seconds in length and contained completely unconstrained spontaneous speech from the 11 different languages. This test set is often referred to as the *45 second utterances* of NIST's 1994 test set. NIST also created a second test set by extracting 614 10-second segments of speech from the original 187 45 second utterances. This test set is referred to as the *10 second utterances*. Results using both test sets will be reported here.

There are many ways to measure the performance of an LID system, including its accuracy and computational efficiency. Computational efficiency is often difficult to compare across systems, since it depends on the specific implementation and computing platform. Therefore, we will focus only on our system's language identification accuracy, as measured by its top-choice accuracy and the rank order statistics. The latter statistic measures the average rank of the correct language within the list of 11, which is indicative of how far down the correct language is from the top-choice answer.

In keeping with House and Neuburg's initial findings, we begin our analysis by focusing on the performance of the system using only the language model for language classification. House and Neuburg's approach suggested that broad phonetic analysis would be less error-prone than detailed classification while still providing reliable LID performance (House and Neuburg, 1977). However, our experiments showed that, despite poor recognition rates, detailed phonetic class representations provide more information than broad classes, and hence yield higher LID accuracy (Hazen, 1993; Hazen and Zue, 1993). Muthusamy *et al.* also concluded that fine phonetic classes were superior for language identification (Muthusamy *et al*, 1993). Experiments have also shown that the exact choice of phones used in the set of fine phonetic units is not critical. We have observed empirically that our LID system achieved only a modest performance improvement when the inventory of phone units was increased to 87 from 59, the number of units used when the phonetic recognizer was trained on English utterances from the NTIMIT corpus (Hazen and Zue, 1994).

Because the language model is the primary component of the system, we decided to investigate in greater detail how the complexity of this component affects performance. Table 2 shows the language identification performance of a unigram, interpolated bigram, and interpolated trigram language model on the NIST test set. Judging from the trend shown in this table, it is conceivable that a further increase in accuracy could be realized by using a higher order n-gram. However, the memory requirement of storing these n-grams for the set of 87 different phones would be prohibitive. An earlier study (Hazen and Zue, 1993) has shown that decreasing the number of phonetic classes helps improve the language identification accuracy of standard (non-interpolated) n-gram models when n is increased. This suggests that the investigation of interpolated class n-grams should be the next step towards improving upon the language model.

To assess the contributions made by each of the individual components, we also measured system performance under conditions in which only one of the components is used at a time as well as when the system utilizes all four components. The results are shown in Table 3. At first glance, it would appear that most of the performance gain in the overall system is contributed by the language model. Closer examination, however, reveals that the contribution made by each individual component depends highly on the length of the utterances. This is shown in Figure 4, in which the top-choice accuracy is plotted against utterance length. As can be seen in these figures, the acoustic model outperforms the language model for shorter utterances. As test utterances get longer, the performance of the language model eventually surpasses and greatly exceeds the performance of the other models. For utterances of 10 seconds or longer, as is the case with all the utterances in the official test set, the language model alone can achieve a performance comparable to that of the complete system. However, this figure shows that additional information beyond the phonotactic information can be useful for increasing language identification accuracy, especially when the utterance is short (< 10 seconds).

As indicated in Table 3, the overall system achieved a top-choice accuracy of 78.1% on the NIST test set, with a rank order statistic of 1.43. For comparison, the best results to date are achieved by systems utilizing Zissman's basic design, which uses a bank of language-dependent phonetic recognizers instead of a single language-independent recognizer. On the same task Zissman's system achieved a top-choice accuracy of 88.8% (Zissman, 1995). However, Zissman's baseline system, which utilizes only phonotactic information (making it comparable to the language model component of our system), achieved an accuracy 79.7%. Because our language model accuracy was 77.5% (see Table 3), this suggests that our system's performance is quite competitive. Zissman's improvement to 88.8% can be attributed to the addition of two new components: (1) gender specific phonetic recognizers, and (2) the incorporation of

quantized duration information directly into the phonotactic model. Yan and Barnard made further modeling and classification refinements to Zissman's basic design to achieve an accuracy of 90.8% (Yan and Barnard, 1995). We believe similar refinements to our modeling techniques should result in performance improvements on the order of those encountered by Zissman and by Yan and Barnard.

In examining the overall performance of the system it is important to examine how the system performs on each individual test language. Table 4 shows the performance of the system as broken down by language. The table also shows the amount of training data available for each language as well as whether or not time aligned phonetic transcriptions were available for any of the training data in each language. As seen in the table, the availability of transcriptions for a particular language appears positively correlated with the system's accuracy for that language. The language-independent phonetic recognizer used by our system is only trained on data from the 6 languages which have transcriptions available. Thus, if a particular language utilizes a phone that is not utilized by one of the six languages used to train the phonetic recognizer, than that phone will never be identified correctly within the hypothesized string of phones generated by the phonetic recognizer.

As seen in Table 4, the performance on Hindi and English utterances is significantly superior to the system's performance at identifying utterances from any other language. This result seems likely to be related to the fact that far more training data was available for training the models of these languages than the other languages. Hindi and English also had transcriptions available for portions of their training sets thus allowing all of their phones to be trained and utilized within the language-independent phonetic recognizer. Likewise, the two languages which performed the worst were Korean and Vietnamese. Neither of these two languages had transcriptions available to be used in training the phonetic recognizer. The fact that both languages contained phones not represented within the multi-language phonetic recognizer, as well as the fact that the two languages had amongst the smallest amounts of training data, probably contributed to the poor performance on these languages. Of the remaining languages, French, Farsi, and Tamil also did not have transcriptions available but did have performances significantly higher than either Korean or Vietnamese. This is probably due in part to the fact that the phone sets of these languages were sufficiently covered by the languages which did possess transcriptions. These results indicate the need for ample, transcribed data from as many of the languages of interest as possible in order to perform highly accurate language identification.

5 Summary

In this paper, we describe a segment-based language identification system motivated by the ideas proposed by House and Neuburg. We formulated the problem into a probabilistic framework, reducing it to four separate components. The system was implemented with a single language-independent phonetic front-end for all languages. This novel approach differs from the most common approaches which utilize multiple single-language phonetic recognizers in the front-end. Using only one phonetic front-end allows the system to be more computationally efficient than the approaches using multiple recognizers in the front end. Additionally, when using phonetic information only and gender independent phonetic recognition, the system is competitive with other state of the art systems. Although the system can achieve good performance based on the phonological language model alone, as suggested by House and Neuburg, other sources of information provide additional performance gain, especially when the system is tested on short utterances.

Aknowledgements

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.

References

[1] R. Cole, M. Fanty, M. Noel, and T. Lander. Telephone speech corpus development at CSLU. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 1815–1818, 1994.

[2] G. Flammia, J. Glass, M. Phillips, J. Polifroni, S. Seneff, and V. Zue. Porting the bilingual VOYAGER system to Italian. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 911–914, September 1994.

[3] J. Glass, D. Goodine, M. Phillips, S. Sakai, S. Seneff, and V. Zue. A bilingual VOYAGER system. In *Proceedings of the 1993 European Conference on Speech Communication and Technology*, pages 2063–2066, September 1993.

[4] T. J. Hazen. Automatic language identification using a segment-based approach. Master's thesis, Massachusetts Institute of Technology, August 1993.

[5] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 1307–1310, 1993.

[6] T. J. Hazen and V. W. Zue. Recent improvements in an approach to segment-based automatic language identification. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 1883–1886, 1994.

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

[8] S. Itahashi, J. X. Zhou, and K. Tanaka. Spoken language discrimination using speech fundamental frequency. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 1899–1902, 1994.

[9] 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, San Mateo, CA, 1990.

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

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

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

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

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

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

[16] 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.

[17] Y. Muthusamy, K. Berkling, T. Arai, R. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 1307–1310, 1993.

[18] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 1007–1010, 1992.

[19] Y. K. Muthusamy, R. A. Cole, and M. 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, 1991.

[20] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 895–898, 1992.

[21] 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, 1983.

[22] Y. Yan and E. Barnard. An approach to language identification with enhanced language model. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, pages 1351–1354, 1995.

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

[24] M. A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1), January 1996.

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

[26] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. Recent progress on the SUMMIT system. In *Proceedings of the Third DARPA Speech and Natural Language Workshop*, pages 380–384, June 1990.

[27] V. Zue, J. Glass, M. Phillips, and S. Seneff. The MIT SUMMIT speech recognition system: A progress report. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 179–189, February 1989.

Monothong Vowels (16)	[i], [ɪ], [ɨ], [e], [æ], [ɛ], [a], [ɑ], [ʌ], [ɔ], [o], [ʏ], [ʊ], [u], [ɝ], [ɯ$^{\text{ɪ}}$]
Diphthongs (6)	[ei], [oʊ], [ɔɪ], [aʊ], [aɪ], [ai]
Semivowels (5)	[w], [j], [y], [l̩], [ɻ]
Flaps & Taps (4)	[ɾ], [ɾ̃], [ɾ$^{\text{ɪ}}$], [ɽ]
Nasals (4)	[m], [n̩], [n], [ŋ]
Fricatives (14)	[β], [f], [v], [θ], [ð], [s], [ʂ], [z], [š], [ç], [x], [ʁ], [h], [ɦ]
Affricates (4)	[ts], [tʃ], [dʒ], [cç]
Stops (12)	[b], [p], [p$^{\text{h}}$], [d], [d̪], [d$^{\text{ɪ}}$], [t$^{\text{h}}$], [t̪], [t$^{\text{ɪ}}$], [g], [k], [k$^{\text{h}}$]
Closures (11)	[b$^{\square}$], [p$^{\square}$], [d$^{\square}$], [d̪$^{\square}$], [d$^{\text{ɪ}\square}$], [t$^{\square}$], [t̪$^{\square}$], [t$^{\text{ɪ}\square}$], [c$^{\square}$], [g$^{\square}$], [k$^{\square}$]
Non-phonetic Units (11)	*background noise*, *filled pause*, *pause*, *breath noise*, *line noise*, *non-speech*, *post-vocalic glottalization*, *onset glottalization*, *lip smack*, *unintelligible speech*, *epinthetic closure*

Table 1: List of phones and non-linguistic descriptors most accuractely describing the 87 phonetic units used within the language-independent phonetic recognizer.

n	Lang. ID Accuracy (%)	Rank Order Statistic
1	68.5	1.73
2	74.3	1.49
3	77.5	1.44

Table 2: Language ID performance of the interpolated n-gram language model using varying n on the 45 second utterances.

Set of Models	10 Second Utterances		45 Second Utterances	
	ACC	ROS	ACC	ROS
Complete System	65.3%	1.83	78.1%	1.43
Language Model	62.7%	1.90	77.5%	1.44
Acoustic Model	49.0%	2.70	53.5%	2.43
Duration Model	31.7%	3.51	44.4%	3.00
F0 Model	12.4%	5.31	20.9%	4.05

Table 3: Performance of complete system and individual components on NIST 1994 test sets using language identification accuracy ('ACC' in table) and rank order statistic ('ROS' in table).

Language	Test Set		# of Training Utterances	Transcriptions Available?
	10 s	45 s		
Hindi	88%	95%	797	Yes
English	87%	89%	1021	Yes
Tamil	70%	71%	525	No
Farsi	67%	79%	438	No
Mandarin	67%	76%	481	Yes
Japanese	66%	79%	408	Yes
French	65%	82%	473	No
German	63%	89%	488	Yes
Spanish	60%	71%	509	Yes
Vietnamese	34%	60%	443	No
Korean	33%	50%	404	No

Table 4: Relationship of the language identification accuracy of the system for particular languages to the size of the language's training set and the availability of transcribed data for that language.

Figure Captions:

Figure 1 System architecture

Figure 2 Illustration of the language identification component of the system

Figure 3 Weighting factors of models over varying utterance length

Figure 4 Performance of system components over varying utterance length as tested on the 1994 NIST 45 second utterance test set

Figure 1:

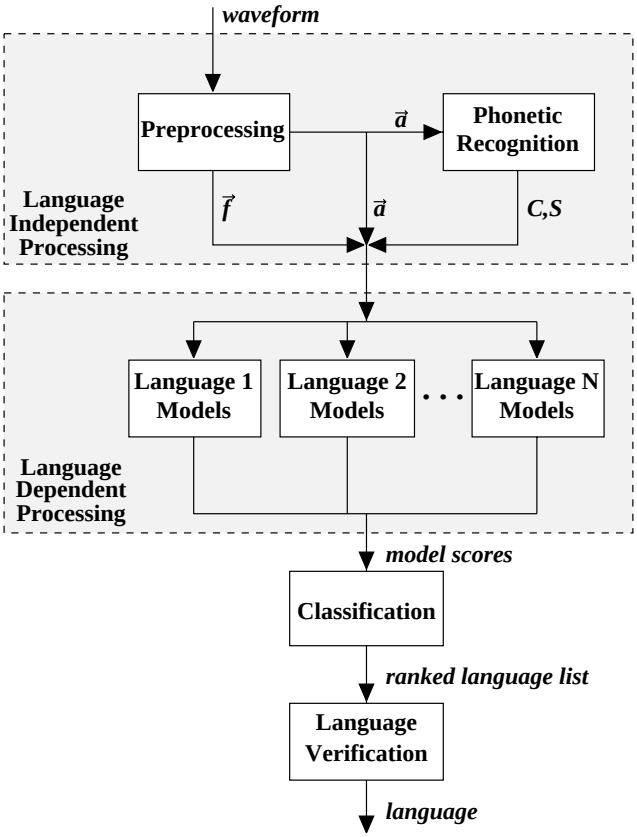

waveform
Preprocessing
$\vec{a}$
Phonetic Recognition
Language Independent Processing
$\vec{f}$
$\vec{a}$
C,S
Language 1 Models
Language 2 Models
Language N Models
Language Dependent Processing
model scores
Classification
ranked language list
Language Verification
language

Figure 2:

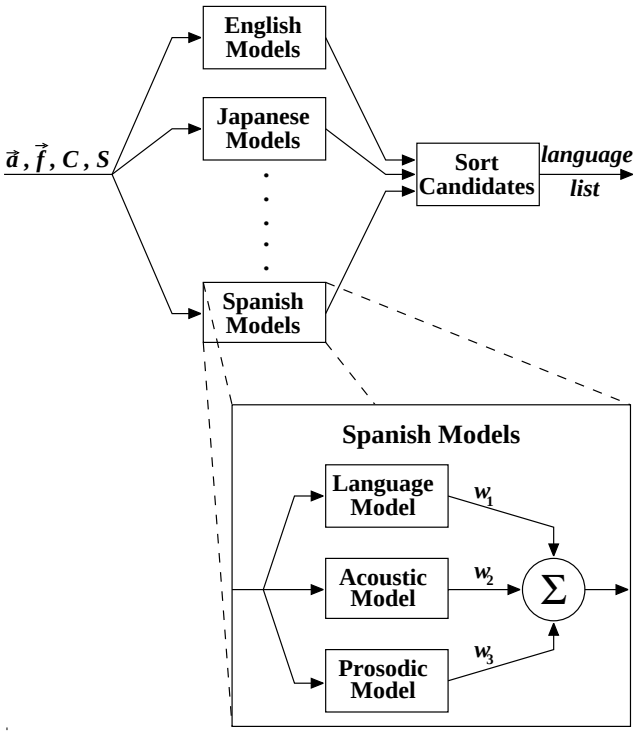

English Models
Japanese Models
Spanish Models
$\vec{a}, \vec{f}, C, S$
Sort Candidates
language list
Spanish Models
Language Model
Acoustic Model
Prosodic Model
w_1
w_2
w_3
Σ

Figure 3:

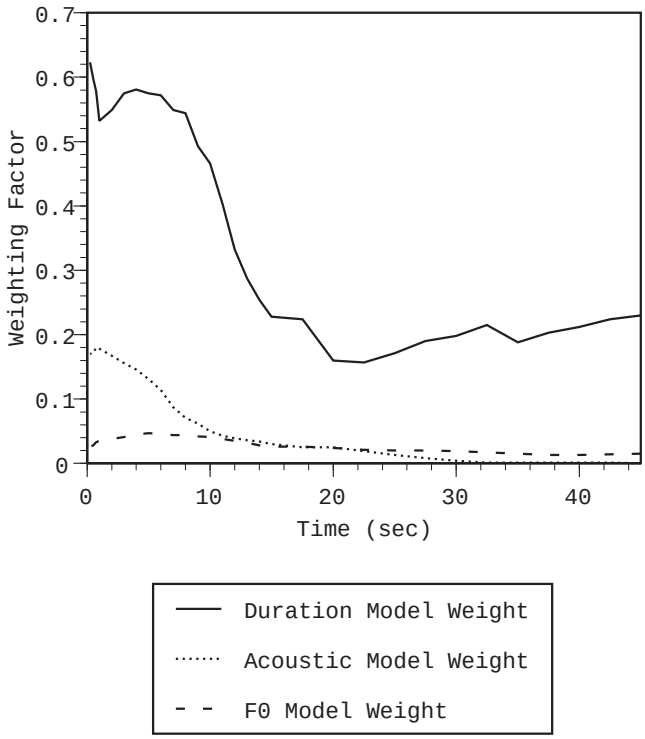
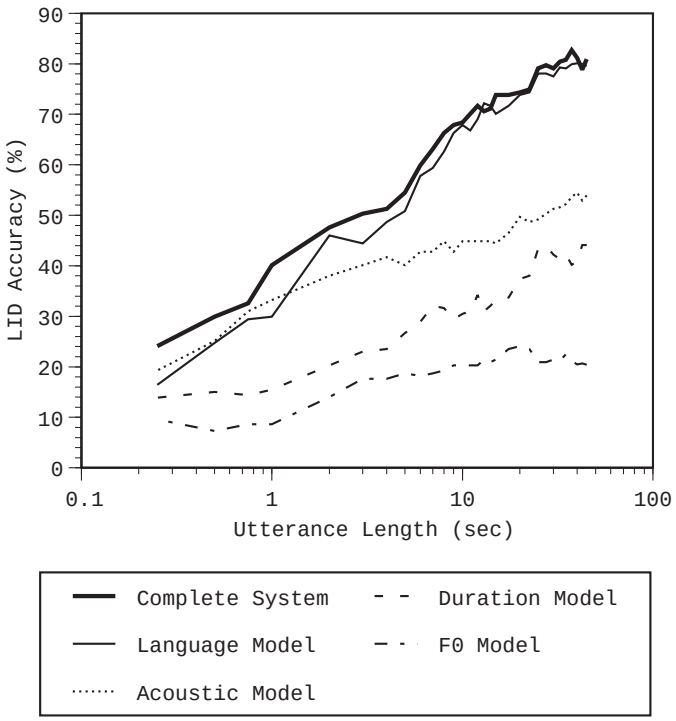

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Weighting Factor
0
10
20
30
40
Time (sec)
Duration Model Weight
Acoustic Model Weight
F0 Model Weight

Figure 4:


LID Accuracy (%)
90
80
70
60
50
40
30
20
10
0
0.1
1
10
100
Utterance Length (sec)
Complete System
Language Model
Acoustic Model
Duration Model
F0 Model