AUTOMATIC LANGUAGE IDENTIFICATION USING A SEGMENT-BASED APPROACH[1]

Timothy J. Hazen and Victor W. Zue

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

ABSTRACT

A segment-based Automatic Language Identification (ALI) system has been developed. The system was designed around a formal probabilistic framework. This framework forms the basis for investigating the ALI approach proposed by House and Neuburg which utilizes phonotactic constraints of languages. The system incorporates different components which model the phonotactic, prosodic, and acoustic properties of the different languages used in the system. The system was trained and tested using the OGI Multi-Language Telephone Speech Corpus. An overall system performance of 47.7% was achieved in identifying the language of test utterances.
Keywords: Automatic language identification.

INTRODUCTION

With recent increased research activities in multi-lingual spoken language systems, interest in Automatic Language Identification (ALI) has grown. The goal of an ALI system is to accurately and efficiently determine the language of a spoken utterance. To date a majority of the research in ALI has focused on frame-based statistical approaches which are trained in an unsupervised manner [1, 2, 3, 4, 5]. While some of these approaches have performed quite well, the work of House and Neuburg suggests that an approach which models the phonotactic constraints of languages could prove extremely effective [6]. Specifically, they proposed that languages can be differentiated strictly based on sequential constraints on phonemes. In fact, the constraints are so strong that they can be captured even if phonemes are described in terms of broad phonetic classes. To provide empirical evidence to their claim, they hand-labeled a corpus of sentences, and showed that high performance language identification can be achieved based on the resulting broad phonetic strings. House and Neuburg's work provides a strong base upon which a segment-based approach to ALI can be built. Despite their compelling demonstration, albeit on hand-labeled data without error, only a few studies have utilized the ideas they introduced [7, 8, 9].

In this paper a segment-based approach to ALI is presented. To provide a solid theoretical foundation for the approach, a probabilistic framework which builds upon House and Neuburg's ideas is first formulated. Based on this framework, a system that incorporates separate models to capture the phonotactic, prosodic, and acoustic information of each language is developed. While our primary goal is to understand the relative merits of each model as system parameters are varied, we will nevertheless measure overall system performance using a publicly available multi-language corpus. Interested readers are referred to a more detailed description of this work in [10].

[1]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.

PROBABILISTIC FRAMEWORK

Before designing the segment-based ALI system, a probabilistic framework describing the ALI problem was derived. To begin, let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ represent the language set of n different languages. An ALI system's basic objective is to determine which of the n languages in $\mathbf{L}$ was used for a particular spoken utterance. The acoustic information of the utterance will be represented by two sets of feature vectors. Let $\vec{\mathbf{a}} = \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$ represent the frame-based set of m feature vectors which encodes the wide-band spectral information of the utterance. Let $\vec{\mathbf{f}} = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_m\}$ represent the frame-based set of m vectors which represent the voicing and fundamental frequency information of the utterance. With these representations, the probability that an utterance was spoken in language L_i, given the acoustic 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 likelihood approach to the problem is to choose the language which is most likely given the acoustic signal. Viewing this as a maximization process the most likely language can be found using the expression

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

The expression in (1) is the most general expression describing the ALI problem. Because every 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 phonologically motivated elements such as phonemes or broad phonetic classes are the most obvious choices for representing the linguistic sequence, $\mathbf{C}$ will be assumed to contain strings of phonetic elements. In a segment-based approach it is necessary to specify how a particular phonetic string C aligns with the acoustic information. To represent this information let $\mathbf{S}$ represent all possible segmentations of the acoustic input. For a particular string with k phonetic elements, $C = \{c_1, c_2, \ldots, c_k\}$ where each c is a particular phonetic label and $S = \{s_1, s_2, \ldots, s_{k+1}\}$ where each s is a segment boundary. To incorporate the phonetic information into the framework, the maximization process in (1) can

be expanded as

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

This expression can be rewritten as

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

With the tremendously large set of possible segmentations and phonetic sequences it would be impractical to perform the summations in (3) over all S and all C. The required computation can be greatly reduced if only a single S and C are used. If it is assumed that the single best hypotheses for S and C can be found independent of the language, it can be shown that (3) can be reduced to

$$\arg\max_i \Pr(L_i \mid C_b, S_b, \vec{\mathbf{a}}, \vec{\mathbf{f}}). \quad (4)$$

where S_b and C_b are the most likely segmentation and phonetic sequence. The expression in (4) can be shown to be equivalent to

$$\arg\max_i \Pr(\vec{\mathbf{a}} \mid C_b, S_b, \vec{\mathbf{f}}, L_i) \Pr(S_b, \vec{\mathbf{f}} \mid C_b, L_i) \Pr(C_b \mid L_i) \Pr(L_i) \quad (5)$$

The four probability expressions in (5) are considerably easier to model separately than the single probability expression in (4). 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. $\Pr(\vec{\mathbf{a}} \mid C_b, S_b, \vec{\mathbf{f}}, L_i) \rightarrow$ The acoustic model.
2. $\Pr(S_b, \vec{\mathbf{f}} \mid C_b, L_i) \rightarrow$ The prosodic model.
3. $\Pr(C_b \mid L_i) \rightarrow$ The language model.
4. $\Pr(L_i) \rightarrow$ The a-priori language probability.

Within this framework, the language model can be used to capture phonotactic information contained in strings of phonemes or broad phonetic classes. The prosodic model can be used to model the prosodic information available in the fundamental frequency contours and segment durations. The acoustic model can capture the manner in which specific phonemes or broad phonetic classes are produced acoustically. The differences that exist within these models from language to language can thus be exploited for the purpose of language identification. Note that if the utterances are evenly distributed amongst the languages, then $\Pr(L_i)$ is the same for all languages, and thus can be ignored in the maximization process.

SYSTEM DESCRIPTION

Corpus

The ALI system described below was trained and tested using the OGI Multi-Language Telephone Speech Corpus [11]. The OGI database consists of utterances spoken in 10 different languages that were collected over the telephone lines. The ten languages are English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Each language contains utterances from 90 different speakers. The database was divided into three sections; 50 speakers per language for the training set, 20 speakers per language for the development test set, and 20 speakers per language for the final test set. For this paper the reported results were obtained by training on the training set and testing on the development test set. Only the text-independent utterances of the corpus where used in training and testing. The final test set has been set aside for future use and was not used for these experiments. The training set contained a total of 2715 utterances while the development set had 1120 utterances. The utterances are roughly evenly distributed amongst the ten languages although the number of utterances per speaker varies from 2 to 6. The male to female ratio of the speakers is roughly 7 to 3. The average duration of each utterance was 13.4 seconds.

Preprocessing

The spectral information of the utterance was represented with 14 mel-frequency cepstral coefficients (MFCC) and 14 delta cepstral coefficients sampled every 5 ms. The voicing information was extracted from the waveform using the *formant* program in Entropic's ESPS package. The pitch tracker returns an estimated F0 value and probability of voicing score every 5 ms. In an attempt to remove speaker dependencies, the $\log_2$ of each F0 value was taken for each voiced frame, and the mean of $\log_2$F0 over the entire utterance was then subtracted away for each frame. A delta F0 value was then computed for each voiced frame from the transformed F0 sequence.

Determination of Segments and Classes

To obtain a broad phonetic string for an utterance, we must first devise a mechanism to determine how an utterance should be segmented, and how these segments can be characterized by a set of broad phonetic classes. In other words, one must develop a phonetic recognizer. Because the OGI corpus was unlabeled at the time of our experiments, the phonetic recognizer could not be trained in a fully supervised fashion. To circumvent this problem, two alternatives were investigated. The first option was to train the recognizer in an unsupervised manner. For this option, all of the utterances were automatically segmented using an adaptation of Glass's multi-level acoustic segmentation algorithm [12]. A threshold was used to obtain a single segmentation from a dendrogram of possible segmentations. For each segment a feature vector of 14 MFCC coefficients averaged over the length of the segment was created. From the set of all segment-based feature vectors in the training set, a codebook was generated using the k-means clustering algorithm. It was our hope that each codeword would roughly correspond to a broad phonetic class (i.e., fricative, vowel, nasal, etc.). However, we found empirically that this assumption may not have been entirely appropriate.

The second option was to train a phonetic recognizer in a supervised manner using other corpora that have been labeled. Since such corpora do not exist for all the languages of interest, we were forced to make a simplifying assumption: While languages differ in their detailed phonetic realizations, the broad phonetic characteristics are nevertheless quite similar. Therefore, we may be able to determine the segments and phonetic classes using a recognizer trained on data from one language. In our case, we used the SUMMIT phonetic recognizer [13, 14] trained on the NTIMIT corpus. The recognizer was then applied to all of the utterances in the OGI corpus to provide the best transcription of English phones for each utterance.

The detailed phonetic labels produced by SUMMIT were then collapsed into broad phonetic classes. Specifically, phone labels with the most similar left and right contexts in the training set were clustered in a hierarchical manner. The hierarchical

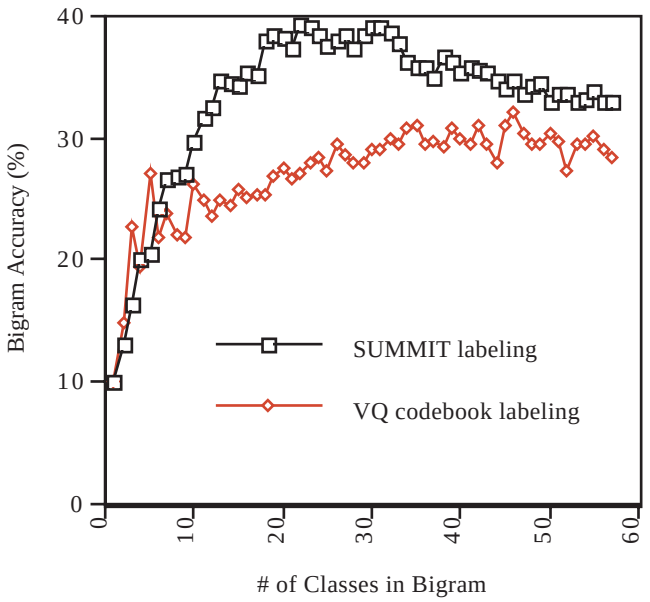

Figure 1: Bigram performance

clustering allowed the number of broad phonetic classes to be varied for our experiments. The similarity measure used for the clustering was the divergence of the probability distribution describing all possible left and right contexts for each phone. Additionally, we also experimented with a few sets of manually selected broad phonetic classes.

Language Modeling

The language model part of (5), $\Pr(C_b \mid L_i)$, can be modeled simply with an n-gram model. Figure 1 shows the performance of the bigram model for the two different phonetic recognizers as the number of broad phonetic classes is varied from 2 to 58. This figure shows that the best performance using the automatically selected broad phonetic classes from the SUMMIT phonetic recognizer was a language identification accuracy of 39.5% (with 22 phonetic classes). Although not shown in this figure, a slight improvement in performance was observed (41.5% with 23 phonetic classes) when the broad phonetic classes were manually selected. In contrast, the best performance using the unsupervised VQ recognizer was 32.1% accuracy with 47 codewords. Our experiments indicate that the bigram model consistently outperformed the unigram and trigram for sequences of broad phonetic classes. This is presumably due to the fact that the bigram offers more constraints than the unigram, and its probabilities can be estimated more reliably than the trigram. For example, the best performance using a trigram model was 34.5% accuracy with 7 manually selected broad classes.

Prosodic Modeling

In (5) the quantity $\Pr(S_b, \vec{\mathbf{f}} \mid C_b, L_i)$ represents the prosodic model. If we assume that the fundamental frequency contours are independent of the phone durations and the phonetic string, then the prosodic model can be rewritten as

$$\Pr(\vec{\mathbf{f}} \mid S_b, C_b, L_i) \Pr(S_b \mid C_b, L_i) \tag{6}$$

and further as

$$\Pr(\vec{\mathbf{f}} \mid L_i) \Pr(S_b \mid C_b, L_i) \tag{7}$$

The expression $\Pr(S_b \mid C_b, L_i)$ can represent a segment duration model. For our experiments, we further assume that the segments are independent of one another. Probability distributions were created to model the number of frames within a segment for each phonetic class of each language. With the 23 broad classes used for the best bigram model, the duration model alone achieved an accuracy of 25.8% in identifying the language of an utterance.

The expression $\Pr(\vec{\mathbf{f}} \mid L_i)$ represents a fundamental frequency contour model. For our experiments, the frames were assumed to be independent. Additionally the F0 and delta F0 measurements were assumed independent. The F0 and delta F0 measurements for all voiced frames were scalar quantized into 150 bins and probability distributions were created for F0 and delta F0 for each language. Using this approach, the F0 model achieved an accuracy of 18.6% in identifying the language of utterances. The delta F0 model achieved an accuracy of 19.7%.

Acoustic Model

The acoustic model $\Pr(\vec{\mathbf{a}} \mid C_b, S_b, \vec{\mathbf{f}}, L_i)$ can be represented with a continuous probability density function which models independent segment-based acoustic feature vectors. For the present experiments the segment-based feature vector consisted of 14 MFCC coefficients and 14 delta MFCC coefficients averaged over all the frames in each segment. To model the feature vector a full covariance Gaussian density function was created for each of the 23 broad classes for each language. The acoustic model alone performed with an accuracy of 33.3%.

System Integration

When the entire set of probabilistic model scores for a single utterance were combined to form one score for each language, it was discovered that the scores for the F0 and delta F0 models were dominating the overall score. Closer examination of each of the individual probabilistic models showed that the top choice probability estimates may have been inflated, as indicated by the fact that the average a posteriori probability for the top choice language was larger than the actual language identification accuracy for each of the models. To compensate for this discrepancy, a scaling factor was applied to the log probability scores of each of the models. For each model, the scaling factors were selected to compress the range of a posteriori probabilities so that the average top choice language probability was equal to the language identification accuracy. This accuracy was obtained from the training data through a jackknifing procedure.

RESULTS

The performance of each model is summarized in Figure 2, together with the overall system performance. Our experimental results suggest that, individually and in descending order, language, acoustic, and prosodic models can achieve some degree of language identification. We view the fact that the language model contributes the most to system performance as supportive of the claim made by House and Neuburg. When all three models are incorporated, the overall system achieved an accuracy of 47.7%. This result is computed by pooling all the utterances in the development set, regardless of length. However, closer examination of the data reveals that the performance improves with the length of the utterance, as shown in Figure 3. The performance of the system improved by almost 10% as the length of the utterances increased from 10 to 45 seconds.

Although the top-choice language identification accuracy was only 47.7%, the correct language was the second and third choice 15.9% and 10.8% of the time, respectively. In other

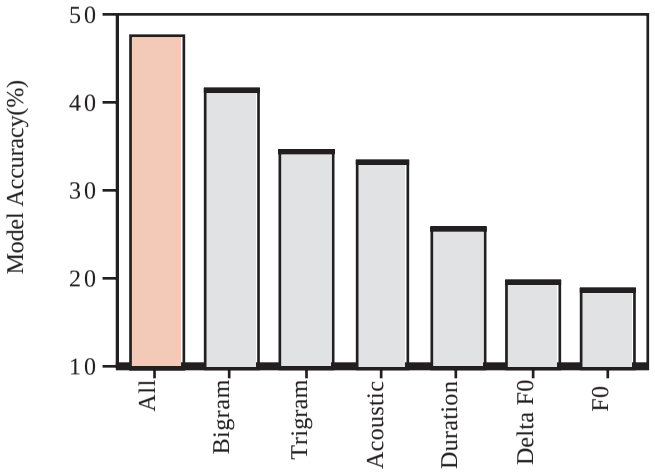

Figure 2: Performance of individual models

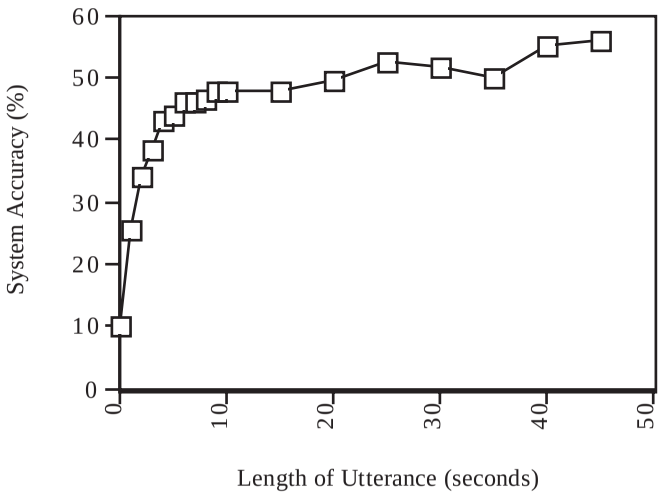

Figure 3: Performance over varying test utterance lengths

words, the correct language was one of the top three choices 74.4% of the time. The rank order statistics, i.e., the average position of the correct language within the ordered list of candidates was 2.65 .

DISCUSSION AND FUTURE WORK

In this paper we formulated the ALI problem within a probabilistic framework. Despite the fact that certain simplifying assumptions were made, the performance of the system that we developed indicates that the phonotactically motivated approach proposed by House and Neuburg can be effective. While our results are very preliminary, they are nevertheless competitive to systems developed by others [5, 9].

Figure 4 show that there is still a large disparity in system performance on training and testing data for 50 training speakers, suggesting that there is plenty of room for performance improvement. We believe that the system will benefit most from improvements to the phonetic recognition and language modeling components of the system. Specifically, future work will include investigating more complex language models, exploring the use of mixture Gaussian models for the acoustic model, and searching for better models to capture the dynamic aspects of the F0 contours.

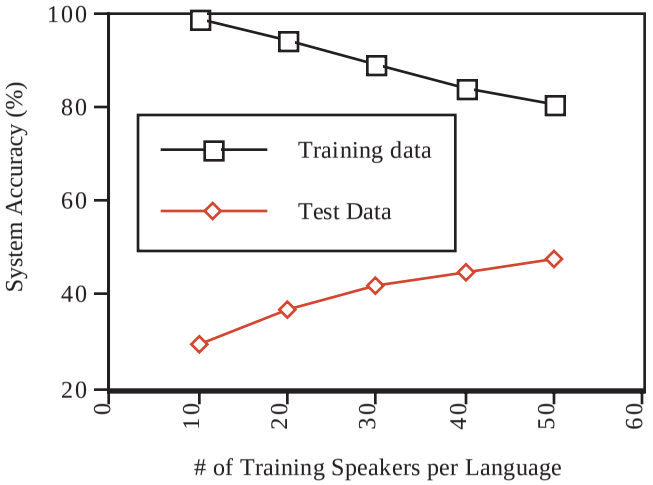

Figure 4: Performance over varying training set sizes

ACKNOWLEDGMENTS

We would like to express our thanks to Ron Cole and Yeshwant Muthusamy of OGI, who generously provided us with the OGI Multi-Language Corpus, and to Mike Phillips of MIT who adapted the SUMMIT system to suit our needs.

REFERENCES

[1] D. Cimarusti and R. B. Ives, "Development of an automatic identification system of spoken languages: Phase I," In *Proc. ICASSP-82,* pp. 1661-1663, 1982.

[2] R. B. Ives, "A minimal rule AI expert system for real-time classification of natural spoken languages," In *Proc. of the 2nd Annual Artificial Intelligence and Advanced Computer Technology Conf.,* pp. 337-340, 1986.

[3] J. T. Foil, "Language identification using noisy speech," In *Proc. ICASSP-86,* pp. 861-864, 1986.

[4] F. J. Goodman, A. F. Martin, and R. E. Wohlford, "Improved automatic language identification in noisy speech," In *Proc. ICASSP-89,* pp. 528-531, 1989.

[5] M. A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," In *Proc. ICASSP-93,* pp. 399-402, 1993.

[6] A. S. House and E. P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *JASA,* Vol. 62, No. 3, pp. 708-713, Sep. 1977.

[7] K. P. Li and T. J. Edwards, "Statistical models for automatic language identification" In *Proc. ICASSP-80,* pp. 884-887, 1980.

[8] Y. K. Muthusamy, R. A. Cole, and M. Gopalakrishnan, "A segment-based approach to automatic language identification," In *Proc. ICASSP-91,* pp. 353-356, 1991.

[9] Y. K. Muthusamy and R. A. Cole, "Automatic segmentation and identification of ten languages using telephone speech," In *Proc. ICSLP 92,* pp. 1007-1010, 1992.

[10] T. J. Hazen, *Automatic Language Identification Using a Segment-Based Approach,* SM thesis, MIT, 1993.

[11] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI Multi-Language Speech Corpus," In *Proc. of ICSLP 92,* pp. 895-898, 1992.

[12] J. R. Glass and V. W. Zue, "Multi-level acoustic segmentation of continuous speech," In *Proc. of ICASSP-88,* pp. 429-432, 1988.

[13] V. Zue, J. Glass, M. Phillips, and S. Seneff, "The MIT SUMMIT Speech Recognition System: A progress report," In *Proc. of the DARPA Speech and Natural Language Workshop,* February, 1989.

[14] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "Recent Progress on the SUMMIT System," In *Proc. of the Third DARPA Speech and Natural Language Workshop,* June, 1990.