RECENT IMPROVEMENTS IN AN APPROACH TO SEGMENT-BASED AUTOMATIC LANGUAGE IDENTIFICATION [1]

Timothy J. Hazen and Victor W. Zue

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA
email: hazen@goldilocks.lcs.mit.edu

ABSTRACT

In 1993, a segment-based system for Automatic Language Identification (ALI) was developed and introduced. The system incorporates phonetic, acoustic, and prosodic information within a probabilistic framework. The original system was trained and tested using the OGI Multi-Language Telephone Speech Corpus and achieved an accuracy of 57.3% in identifying the language of test utterances from the OGI corpus. Recent improvements to the system have included the addition of channel normalization during preprocessing, the utilization of the recently transcribed utterances from the OGI corpus for phonetic recognition training, the use of mixture Gaussian density functions for the modeling of prosodic information, and the development of a hill-climbing optimization procedure for determining the scaling factors used when combining the scores from different models. The current system has achieved an accuracy of 79.7% in identifying the language of test utterances.

INTRODUCTION

Recently, research activities in Automatic Language Identification (ALI) systems have increased in conjunction with the growing interest in multi-lingual spoken language systems. Multi-lingual systems may require an accurate and efficient means for determining the language of a spoken utterance. While higher level information, such as vocabulary and syntactic constraints, may be used to uniquely determine the language of an utterance, utilization of this information for a large set of potential languages could be computationally expensive. However, examination of lower level information, such as phonotactic constraints, may provide enough information to allow for accurate language identification without being computationally burdensome. If a small subset of possible languages can be identified quickly (i.e. via a *fast match*), then higher level information can be used to verify the top-choice language. This paper presents a segment-based ALI system which is intended to provide a *fast match* list of likely candidate languages for a spoken utterance. This system, previously described in [1] and [2], incorporates phonetic, acoustic and prosodic information within a probabilistic framework.

[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.

SYSTEM DESCRIPTION

Corpus

The ALI system described herein was trained and tested using the OGI Multi-Language Telephone Speech Corpus [3]. The original OGI database consisted 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 contained 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. Within the last year the database has been expanded to include additional utterances in the ten original languages as well as utterances in Hindi.

For this paper, tests using several different training and test sets were conducted. The initial system was developed using 2715 utterances from the original training set for training, and tested using 1120 utterances from the development test set. The second set of tests used the original training set, the development test set, and the utterances in the expanded data set for training while the final test set was used for testing. For these experiments 5987 utterances were available for training. Of these, 552 utterances distributed amongst English, German, Hindi, Japanese, Mandarin, and Spanish were accompanied by a full time-aligned phonetic transcription. The test set for these experiments contained 187 utterances as selected by NIST[2] for their March 1994 evaluation. These utterances were a minimum of 30 seconds in length and contained completely unconstrained spontaneous speech from 11 different languages.

[2]The National Institute of Standards and Technology.

System Architecture

The system consists of three primary sub-systems: the preprocessor, the phonetic recognizer, and the language identifier. The preprocessor takes the raw waveform and produces frame-based feature vectors representing the wide-band spectral information and the fundamental frequency (F0) and voicing information. The wide-band spectral information is represented as $\vec{\mathbf{a}}$ where $\vec{\mathbf{a}} = \{\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m\}$ and each $\vec{a}$ is an individual frame. Similarly, the fundamental frequency information is represented as $\vec{\mathbf{f}} = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_m\}$. The phonetic recognizer uses the acoustic information in $\vec{\mathbf{a}}$ and $\vec{\mathbf{f}}$ to produce the string of phonetic elements that are most likely. This string of phonetic elements is represented as C where $C = \{c_1, c_2, \ldots, c_k\}$ and each c is a phonetic element. The phonetic recognizer also provides a segmentation to accompany the most likely string of phones. This segmentation is represented as $S = \{s_1, s_2, \ldots, s_{k+1}\}$ where each s is a segment boundary. Using $\vec{\mathbf{a}}$, $\vec{\mathbf{f}}$, C, and S the language identifier generates an ordered list of language hypotheses from the language set $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$.

Preprocessing

The wide-band spectral information of the utterance is represented with 14 mel-frequency cepstral coefficients (MFCC's) and 14 delta cepstral coefficients sampled every 5 ms. Mean cepstral subtraction is utilized to reduce the effects of the acoustic differences of the different channels present in the data.

The F0 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 $\log_2$F0 value was also computed for each voiced frame.

Phonetic Recognition

The SUMMIT phonetic recognizer [4, 5] is used to determine the most likely string of phonetic elements C and segmentation S. The phonetic string C is represented using 87 different phones. The recognition phase is completely language independent. The recognizer is trained using the phonetically transcribed utterances in the OGI corpus.

Probabilistic Framework

Given the outputs $\vec{\mathbf{a}}$, $\vec{\mathbf{f}}$, C, and S from the preprocessor and the phonetic recognizer, the language identification problem can be approached as a maximum *a posteriori* probability problem. In this case, the top choice language of the language identifier will be the language which is most likely given the values of $\vec{\mathbf{a}}$, $\vec{\mathbf{f}}$, C, and S. This can be expressed as

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

This expression can be shown to be equivalent to

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

The four probability expressions in (2) are considerably easier to model separately than the single probability expression in (1). Additionally, the expression is now organized in such a way that prosodic, phonotactic and acoustic information are contained in separate terms. In modeling, these terms will be referred to as:

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

Within this framework, the phonological language model can be used to capture phonotactic information contained in strings of 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 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. For the experiments in this paper, it is assumed $\Pr(L_i)$ is the same for all languages, and thus can be ignored.

Language Model

The language model part of (2), $\Pr(C \mid L_i)$, can be modeled simply as an n-gram. The experiments used in this paper use an interpolated trigram model which can be expressed as

$$\Pr(C \mid L_i) = \prod_k \hat{P}(c_k \mid c_{k-1}, c_{k-2}, L_i). \tag{3}$$

where

$$\hat{P}(c_k \mid c_{k-1}, c_{k-2}, L_i) = \lambda_1 \Pr(c_k \mid c_{k-1}, c_{k-2}, L_i) + (1 - \lambda_1) \hat{P}(c_k \mid c_{k-1}, L_i) \tag{4}$$

and

$$\hat{P}(c_k \mid c_{k-1}, L_i) = \lambda_2 \Pr(c_k \mid c_{k-1}, L_i) + (1 - \lambda_2) \hat{P}(c_k, L_i). \tag{5}$$

The λ values must be chosen such that the trigram score dominates the interpolated score when sufficient training data exists to properly estimate the trigram probabilities but carries less weight when an insufficient amount of training data exists. Thus λ_1 and λ_2 are defined as

$$\lambda_1 = \frac{k_{c_{i-1}, c_{i-2}}}{k_{c_{i-1}, c_{i-2}} + K_1}, \text{ and } \lambda_2 = \frac{k_{c_{i-1}}}{k_{c_{i-1}} + K_2} \tag{6}$$

where K_1 and K_2 are constants, and $k_{c_{i-1},c_{i-2}}$ and $k_{c_{i-1}}$ are the counts of the number of times the respective bigram and unigram substrings occurred in the training data. The constant values K_1 and K_2 were empirically set to the values of 850 and 300 based on jackknifing experiments designed to optimize the performance of the language model on data from the training set.

Acoustic Model

The acoustic model part of (2), $\Pr(\vec{\mathbf{a}} \mid C, S, \vec{\mathbf{f}}, L_i)$ can be simplified if segment independence is assumed. The acoustic model can thus be expressed as

$$\Pr(\vec{\mathbf{a}} \mid C, S, \vec{\mathbf{f}}, L_i) = \prod_k \Pr(\vec{\alpha}_k \mid c_k, L_i) \tag{7}$$

where each α_k is a segment-based feature vector. The feature vectors for these experiments consisted of 14 MFCC's and 14 delta MFCC's averaged over all the frames in each segment. To model the feature vector, mixtures of diagonal Gaussian density functions were utilized for each of the 87 phonetic classes in each language.

Prosodic Model

In (2) the quantity $\Pr(S, \vec{\mathbf{f}} \mid C, L_i)$ represents the prosodic model. In the original system, the fundamental frequency contours are assumed to be independent of the phone durations and the phonetic string. With this assumption the prosodic model is expressed as

$$\Pr(\vec{\mathbf{f}} \mid L_i) \Pr(S \mid C, L_i). \tag{8}$$

The expression $\Pr(S \mid C, L_i)$ will be referred to as the segment duration model. For our experiments, we further assume that the segments are independent of one another, allowing the duration model to be represented as

$$\Pr(S \mid C, L_i) = \prod_k \Pr(d_k \mid c_k, L_i) \tag{9}$$

where d represents the duration of a segment. The duration of segments for each phonetic class in each language are modeled with mixtures of Gaussian density functions.

The expression $\Pr(\vec{\mathbf{f}} \mid L_i)$ represents an F0 contour model. For our experiments, the frames are assumed to be independent, allowing the F0 model to be modeled as

$$\Pr(\vec{\mathbf{f}} \mid L_i) = \prod_m \Pr(\vec{f}_m \mid L_i) \tag{10}$$

where only voiced frames are used in finding the model's score. The frame-based feature vectors are modeled with mixtures of full covariance Gaussian density functions for each language.

The complete system used in the development tests utilized the two separate duration and F0 models described above. However, because segment durations and fundamental frequency contours may jointly contribute to the prosodic composition of an utterance, it may not be appropriate to assume they are independent. An alternate approach which accounts for the within-segment correlation of duration and F0 features can be expressed as

$$\Pr(S, \vec{\mathbf{f}} \mid C, L_i) = \prod_k \Pr(\vec{p}_k \mid c_k, L_i) \tag{11}$$

where each $\vec{p}$ is a segment-based prosodic feature vector which includes the duration, average F0 and average ΔF0 for voiced segments and only the duration for unvoiced segments. For this approach, mixtures of diagonal Gaussian density functions are used to model the segment-based prosodic feature vectors for each phonetic class in each language.

System Integration

In principle, the log probability scores from each individual model need only be added to generate the full system score for a particular language given a test utterance. Unfortunately, in reality the acoustic and F0 models dominate the full system score, completely obliterating the effect of the language model. To compensate for this the log probability score for each model can be multiplied by a scaling factor. Two methods for selecting the scaling factor have been investigated. Both methods involve optimizing the scaling factors based on tests performed on development test data jackknifed from the training set.

Closer examination of each of the individual probabilistic models shows that the top choice probability estimates appear inflated, as indicated by the fact that the average *a posteriori* probability for the top choice language is larger than the actual language identification accuracy for each of the models. To compensate for this discrepancy, scaling factors can be applied to the log probability scores of each model where the scaling factors are selected to compress the range of *a posteriori* probabilities for that model so that its average top choice language probability equals its language identification accuracy. This method was used in all of the development tests.

The second method is to optimize the scaling factors so that the full system achieves a maximal language identification accuracy on development tests. This is accomplished with a hill-climbing search routine where the system's scaling factors are optimized one model at a time in an iterative fashion until a local maximum in performance is achieved.

RESULTS

Development Results

As shown in the first row of Table 1, our first implementation of an ALI system, as reported in [1], achieved an accuracy of 47.7% when tested on the complete de-

Date	Comments	Accuracy
4/93	System presented in [1]	47.7%
8/93	System presented in [2]	48.6%
1/94	+ Channel normalization	54.8%
1/94	+ Mixture Gaussian duration model	55.8%
2/94	+ Recognizer trained w/ OGI data	58.5%

Table 1: Summary of development test results

Test date	March '94		June '94	
Utt. length	10 sec.	>30 sec.	10 sec.	>30 sec.
Language model	61.6%	72.7%	62.7%	77.5%
Acoustic model	48.8%	52.9%	49.0%	50.8%
Duration model	34.7%	43.3%	31.7%	44.4%
F0 model	12.4%	20.9%	12.4%	20.9%
Complete system	65.4%	70.1%	62.6%	69.0%

Table 2: Summary of recent test results on NIST's March '94 evaluation test set (all values are accuracy percentages)

velopment test set.[3] The system described in [2] contained incremental improvements over the system in [1] and, as shown in the second row of Table 1, achieved an accuracy of 48.6% on the development test set. Since then significant improvements have been made to the system. These improvements include (1) the addition of channel normalization during preprocessing, (2) the use of mixture Gaussian modeling in the duration model as opposed to non-parametric modeling techniques, and (3) the use of a phonetic recognizer trained on the phonetically transcribed utterances from the OGI corpus as opposed to one trained on the NTIMIT corpus (which contains only English read speech). The remaining rows of Table 1 summarize the contributions each of these modeling techniques has made in improving the performance of the system on the development test set.

Recent Experiments

Table 2 shows the results of our current system for two separate tests using NIST's March '94 evaluation test set. The two different experiments both utilize the *a posteriori* probability adjustment method for determining the scaling factors for combining the separate models. The only difference between the March '94 test and the June '94 test is the number of utterances used to train the phonetic recognizer. The March '94 test only used the 297 transcribed utterances within the original training set. The June '94 test used the 552 transcribed utterances described previously.

Table 3 compares the performance of the system as the means for determining the model scale factors is changed from the *a posteriori* probability adjustment method to the hill-climbing optimization method. When the hill-climbing optimization method is utilized the performance increases from 69.0% to 79.7%. Table 3 also shows the system's performance when a single prosodic model is used in place of the duration and F0 models.

[3]The same system achieved an accuracy of 57.3% on the 178 utterances of NIST's April '93 evaluation set, which also contained unconstrained utterances of length 30 seconds or greater.

Test date	June '94	
Utt. length	10 sec.	>30 sec.
Complete system	62.6%	69.0%
+ hill climbing optimization	65.9%	79.7%
+ combined prosodic model	65.0%	79.1%

Table 3: System performance with new modeling techniques added (all values are accuracy percentages)

DISCUSSION

There are several key observations that can be made from the results. First, channel normalization preprocessing greatly improves the performance of the system. Second, an improper means of selecting the scaling factors used when combining the individual model scores can hurt the system. The hill-climbing optimization method for selecting the scaling factors helped increase the performance of the system above the level of any of the individual models. However, it is not clear whether this method is the most appropriate method. Last, combining the F0 and duration models into a single prosodic model had little effect on the system's performance, though more investigation may still be required. Overall, the improvements described in this paper have increased the performance of the system from an accuracy rate of 57.3% up to 79.7% over the last year.

ACKNOWLEDGMENTS

We would like to express our thanks to Li Lee for assisting in the development and testing of our system, to Mike Phillips for his help in adapting SUMMIT to fit our needs, and to Jim Glass for developing the mixture Gaussian code used by our system.

REFERENCES

[1] T. J. Hazen and V. W. Zue, "Automatic language identification using a segment-based approach," In *Proc. Eurospeech 93*, pp. 1303-1306, 1993.

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

[3] 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.

[4] 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,* pp. 179-189, February, 1989.

[5] 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,* pp. 380-384, June, 1990.