

ROBUST SPOKEN LANGUAGE IDENTIFICATION USING LARGE VOCABULARY SPEECH RECOGNITION.

*James L Hieronymus and Shubha Kadambe**

Bell Laboratories, 700 Mountain Avenue, Murray Hill, NJ 07974
Atlantic Aerospace Elect. Corp., 6404 Ivy Lane, Greenbelt, MD 20906*

ABSTRACT

A robust, task independent spoken Language Identification (LID) system which uses a Large Vocabulary Continuous Speech Recognition (LVCSR) module for each language to choose the most likely language spoken is described. The acoustic analysis uses mean cepstral removal on mel scale cepstral coefficients to compensate for different input channels. The system has been trained on 5 languages: English, German, Japanese, Mandarin Chinese and Spanish using a subset of the Oregon Graduate Institute 11 language data base. The five language results show 88% correct recognition for 50 second utterances without using confidence measures and 98 % correct with confidence measures without the robust front end. The recognition rate is 81 % correct for 10 second utterances without confidence measures and 93 % correct with confidence measures without the robust front end. Adding the robust front end improves the recognition rate approximately 3 % on the short utterances and 1 % for the long utterances. The best performance has been obtained for systems trained on phonetically hand labeled speech.

1. INTRODUCTION

In the future, Language Identification (LID) systems will be an integral part of telephone and speech input computer networks which provide services in many languages. A LID system can be used to pre-sort the callers into the language they speak, so that the required service will be provided in the language appropriate to the talker. Examples of these services include, travel information, emergency assistance, language interpretation, telephone information, buying services, banking and stock trading. Since there are large numbers of non-English talkers in the US population due to immigration, there is a need to offer multi-language capability even within the US. International markets and tourism add to the desirability of offering services in many languages.

The languages of the world differ from one another along many dimensions which have been codified as linguistic categories. These include, phoneme inventory, phoneme sequences, syllable structure, prosodics, lexical words and grammar. Therefore, we hypothesize that an LID system which exploits each of these linguistic categories in turn will have the necessary discriminative power to provide good performance on short utterances.

This paper is divided into five sections. First, past work is discussed. Second, the basic phoneme recognition system is described. Third, the architecture of the LVCSR system is discussed. Fourth, details of training the system for each language are given. Then, the LID results for the system using LVCSR sub-systems are discussed. These show that the more complete language modeling which the LVCSR system provides gives the best performance. Our original method for computing the probability an acoustic input corresponds to a certain language, neglected the probability of the acoustic sequence within the language which appears in the numerator due to Bayes' Rule. This was estimated and improved the performance greatly. Finally the results of using mean cepstral subtraction to make the system more robust to differences in the telephone channel are presented.

2. PAST WORK

Neuburg and House [2] used an ergodic Markov model of sequences of 5 broad phonetic categories (stop consonant, fricative consonant, non-vocalic sonorant, vowel and silence) to identify 8 languages with 100% accuracy using phonetic hand labels for input, when they tested on the training data (because of the scarcity of data for the experiment). The system was designed to have a variable number of states (from 2 to 5) to model each language. A version of our system also gets 100% correct language identification using automatically obtained fine phonetic labels and a trigram phonemotactic model (corresponding to a 3 state model in the Neuburg and House system) for five languages when tested on training data.

Recently, there has been much interest in phonetic modeling of speech for language identification. Muthusamy et al developed a language identification system based on broad phonetic classes and neural network classifiers [3]. Muthusamy [4] also collected an 11 language telephone speech database at Oregon Graduate Institute which became the standard training and testing database for a series of U.S. Government sponsored language identification tests which were administered by the National Institute for Standards and Technology (NIST). These tests provide a way of measuring the relative performance of many systems incorporating different methods of spoken language identification. The results reported in this paper are obtained from the Spring 1994 training and test data for these tests.

In the past three years, a number of researchers [5, 6, 7] have

been developing systems which first recognize phonemes using HMM phoneme modeling and then use a phonemotactic model of phoneme sequences allowed within each language to identify the spoken language. Our baseline system described in [7] uses Continuous Density second order Variable Duration Hidden Markov Model (CVDHMM) to achieve the phoneme recognition based on context conditioned phonemes (called tri-phones in the literature) and trigram phoneme sequence models (phonemotactic models). The other LID systems [6, 5] mentioned above, use context independent first order Markov models for phoneme recognition with bigram phonemotactic models. For languages with Consonant-Vowel-Consonant (CVC) syllable structure, the trigram models do a very good job of modeling the most frequent words, which are usually monosyllabic and hence, should help in discriminating these languages more efficiently than bigram phonemotactic models. On the other hand languages with Consonant-Vowel syllable structure would be more efficiently modeled by the bigram phonemotactic models. The languages studied represented a mixture of syllable structures, so that the trigram phonemotactic models seem to provide an advantage.

This paper shows that full LVCSR leads to better LID performance, than the same system using trigram phonotactics, and language specific phonemes alone. Another system developed at Dragon Systems [14, 15] also uses full LVCSR for LID and performs very well on three languages. Developing automated methods for creating LVCSR systems for new languages would allow these performance advantages to be realized for language identification without extensive human effort.

3. DESCRIPTION OF THE PHONEME RECOGNITION SYSTEM

The phoneme recognition system was developed at Bell Labs for English by A. Ljolje [1]. This phoneme recognizer is based on a second order ergodic CVDHMM. The ergodic HMM has one state per phoneme. However, the acoustic model for each phoneme is a time sequence of three probability density functions (pdf's) with each pdf representing the beginning, the middle and the end of a phoneme, respectively. The pdf's are represented as mixtures of Gaussian pdf's on the acoustic features which have been de-correlated. This structure is equivalent to a three state left-to-right HMM phoneme model. The duration of each phoneme is modeled by a four parameter gamma distribution function. The four parameters are: (1) the shortest allowed phoneme duration (the gamma distribution shift), (2) the mean duration, (3) the variance of the duration, and (4) the maximum allowed duration for the phoneme. The shortest allowed duration is equal to the shortest observed duration in the training data, the mean and variance are calculated from the training data and the maximum duration is calculated as the 95th percentile of the distribution. Because the ergodic HMM is second order, the transition probabilities are the probability of the next phoneme given the past two phonemes. This is then the trigram phoneme sequence probability which can be estimated separately. A diagram of a Second Order Ergodic HMM is shown in Figure 1.

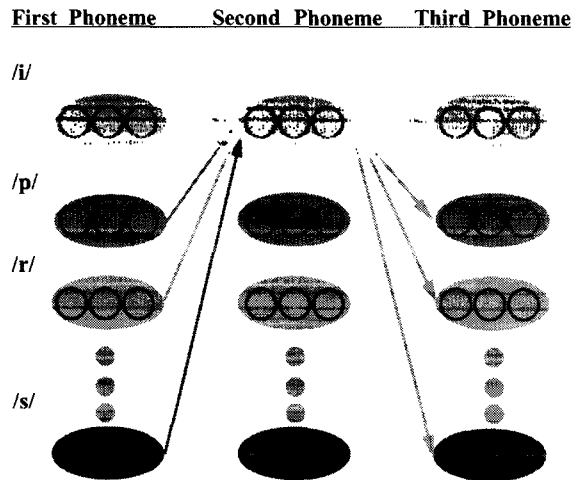


Figure 1. The architecture of a second order ergodic CVDHMM.

3.0.1. Acoustic features

For the baseline system, the speech feature vector consists of 26 features which were chosen from 38 coefficients of 12 cepstra, 12 delta cepstra, 12 delta delta cepstra, delta energy and delta delta energy using a discriminant analysis method. [9] First all the cepstral coefficients are computed using an autocorrelation LPC model with a 20 msec time window which is shifted by 10 msec per frame. Then the coefficients are de-correlated (rotated to be orthogonal). The De-correlation is most needed for the cepstral and delta delta cepstral coefficients.

For the robust system, the speech feature vector consists of 38 mel scale cepstral coefficients, consisting of 12 cepstra, 12 delta cepstra, 12 delta delta cepstra, delta energy and delta delta energy. The means of each cepstral coefficient across the utterance are subtracted from the frame by frame coefficients to normalize for differences in channel characteristics. Then the De-correlation is performed as in the baseline system.

3.0.2. Training phoneme models

The best performance came from training the system with phonetic hand labels which were available for these 5 languages. The initial models are trained using this data and the models are re-trained using the segmental k-means algorithm iteratively until the models converge, in three iterations. We speculated that in spontaneous speech many segments are deleted or severely coarticulated. Training on the orthography and a text to speech system gives many extra segments which have not been realized in the speech, which gives a bad acoustic model for the often deleted phonemes.

4. LID SYSTEM USING LARGE VOCABULARY SPEECH RECOGNITION

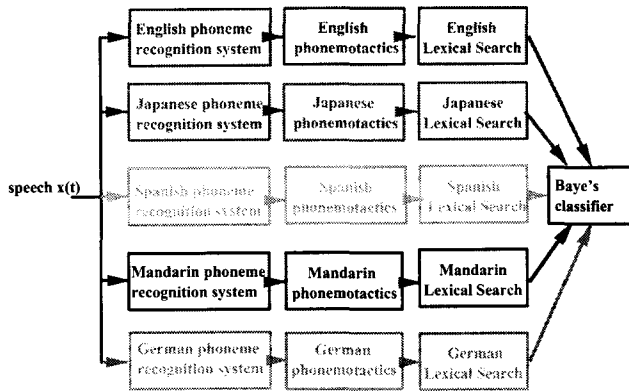


Figure 2. The block diagram of the complete LID system

A full LVCSR system is used for language identification. The block diagram of the LID system is as shown in Figure 2.

The lexical access system uses cascades of weighted finite state transducers [10] to do lexical search and grammatical constraints. The first step is a transduction from phoneme lattices to word lattices. The best path through the finite state network provides the most probable word sequence and the probability, for words in the vocabulary without a word sequence model. The second transduction is from a word lattice to a sentence lattice, which obeys the bigram word grammar trained on the OGI training and development test portions of the corpus. The best path through the resulting sentence lattice is the most probable sentence given the language model. For language identification, the subsystems (block 1, 2 and 3 in Figure 2) for each language are run in parallel for a given speech signal similar to the base line system described above. The language subsystem with the highest normalized log likelihood is chosen as the language of the input speech signal.

5. TRAINING THE LID SYSTEM

The five language (English, German, Mandarin Chinese, Japanese and Spanish) LID system was trained and tested using the prompted monologue section of the 11 language speech data base collected by OGI [4]. The training and test data consists of about 80 and 18 speakers, respectively, per language. The monologue recording is 50 seconds in length, including pauses, which yields 35-45 sec of speech.

The acoustic HMM models systems were trained on phonetically hand labelled speech.

5.1. Training the Word and language model

For each of the 5 languages, the vocabulary was chosen by taking every unique word from the prompted monologues in the training and devtest portion of the OGI 11 language database. Table 1 below shows the number of words in the lexicon for each language and their average length in phonemes. A bigram language model was trained for each language using the Katz backoff method described in [11]. An attempt was made to add text from newspaper sources

to augment the language model, but this resulted in poorer performance, because newspaper style is very different from spoken language.

Language	# words	average length
English	2564	7.47
Spanish	2014	11.36
German	1844	8.34
Japanese	1863	7.80
Mandarin	1546	4.07

Table 1. Lexicon Sizes for Each Language

5.2. Training the Final Classifier

The LVCSR LID system has many differences in the language models. Different languages have a different number of phonemes, different average word lengths, different word sequences which may contain high frequency words like particles and articles. The result of these differences is that the scores from each subsystem has to be normalized relative to the scores for the other languages. Additive and multiplicative factors have been considered (which for log likelihoods correspond to multiplicative and exponential factors on the probabilities). The additive factors seemed to work best, with the factors trained on the home town section of the OGI corpus. The language dependent normalizations were much smaller in the system which normalized by the acoustic probability as discussed below.

6. EXPERIMENTAL DETAILS AND RESULTS

The baseline LVCSR system was tested using 1994 LID evaluation test set. The system identified the language spoken an average of 88 % LID rate on five languages on the 50 secs utterances and 81 % on the 10 secs utterances. In Table 2, the results for five languages are shown.

In order to improve the identification rate for the LVCSR system, a technique which involves normalizing the utterance recognition log likelihood, by the log likelihood of the best unconstrained phoneme sequence has been tried. This is similar to the technique used by Rose and Paul [12] for keyword spotting and by Young and Ward to detect out of vocabulary words in a large vocabulary ASR system [13]. Later publications refer to this as a confidence measure, and is a ratio between the lexicon and grammar constrained best path probability and the unconstrained acoustic model path probability through the utterance. It was first used for LID

Length and cond	English	German	Spanish	Japanese	Mandarin
50 sec no gram	85 %	90 %	76 %	93 %	94 %
50 sec gram	95 %	95 %	94 %	97 %	97 %
50 sec norm	98 %	99 %	96 %	98 %	98 %
50 sec norm robust	99 %	99 %	99 %	99 %	99 %
10 sec no gram	81 %	82 %	79 %	83 %	81 %
10 sec gram	85 %	84 %	81 %	87 %	85 %
10 sec norm	95 %	96 %	90 %	92 %	90 %
10 sec norm robust	98 %	97 %	95 %	97 %	95 %

Table 2. Five language identification results.

by Lowe et al. [14]. Further results are reported in Mendoza et al. [15]. The equation for the recognition of an utterance in a particular language is

$$P(S_i, L_i, P_i|x) = P(x|\beta_i)P(\beta_i|W_i)P(S_i|W_i)/P(x) \quad (1)$$

where the P s are probabilities, x is the input speech signal, β_i is the phoneme sequence, W_i is the word sequence, S_i is the set of all possible sentences for language i and L_i is the phonemotactic model of the language i . The term in the numerator is often considered to be the same for all the of sentences recognized, and thus neglected. This is no longer true when we wish to make a comparison between the output of recognizers for different languages. The estimation of this term gives a better estimate of the total probabilities and thus better performance when comparing different systems. It may also be possible to allow rejection of languages not in the set trained. The method we used to estimate the probability of the acoustic sequence is to run the phoneme recognizer with equal trigram probability for all possible phoneme sequences. A better estimate might be to compute the best acoustic score based on the best match for all of the Gaussian mixtures in the system. This was the measure used by Lowe et al. [14] and will be explored in future work. The present acoustic probability estimate raises the performance of the system to the final best result of 98 % correct identification for 50 sec and 93 % correct identification for 10 sec of speech. The results are shown in Table 2 with the label norm.

7. ROBUST LANGUAGE IDENTIFICATION

The robust system differs only in the front end processing. It produced results which are approximately 3 % better than the baseline system for the short utterances, 99 % correct identification for 50 sec and 96 % correct identification for 10 sec of speech. The results are shown in Table 2 with the label norm robust. This is probably due to the fact that the effect of the different telephone microphones, telephone channels and noise conditions are normalized out by using mean cepstral removal. These techniques have been used in LVCSR in the past with similar improvements in performance. [16] Thus the system appears to be more robust than the baseline system.

8. CONCLUSIONS

A robust five language identification system based on LVCSR was described. LID results for five languages were

best results for the full LVCSR system, with a normalization or confidence estimate based on unconstrained phoneme matches for each language. Mean cepstral subtraction gave an approximately 3 % performance gain over the baseline system for the short utterances.

9. ACKNOWLEDGMENTS

We would like to thank Andrej Ljolje, Mike Riley, Fernando Pereira, Richard Sproat, Bernd Moebius, Chilin Shih and Padma Ramesh for helpful discussions.

REFERENCES

- [1] A. Ljolje, *High Accuracy Phone Recognition Using Context Clustering and Quasi-triphonic Models*, Computer Speech and Language, to appear.
- [2] A. S. House and E. P. Neuberg, (1977) "Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations", *J. Acoust. Soc. Am.*, 62, 708-713, 1977.
- [3] Y. K. Muthusamy and R. A. Cole, (1992), "Automatic segmentation and identification of ten languages using telephone speech," *Proc. of ICSLP 92*, Banff, Canada
- [4] Y. K. Muthusamy, R. A. Cole and B. T. Oshika, (1992), "The OGI Multi-Language Telephone Speech Corpus", *Proc. of ICSLP 92*, Banff, Canada.
- [5] M. A. Zissman and Elliot Singer, (1994) "Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modeling", *Proc. of ICASSP 94*, Adelaide, Australia, April 1994.
- [6] L. F. Lamel and J. L. Gauvain, (1994), "Language Identification Using Phone-based Acoustic Likelihoods", *Proc. of ICASSP 94*, Adelaide, Australia, 1994.
- [7] S. Kadambe and J. L. Hieronymus, (1994, "Spontaneous speech language identification with a knowledge of linguistics", *Proc. of ICSLP 94*, Yokohama, Japan, pp.
- [8] S. Kadambe and J. L. Hieronymus, (1995, "Language Identification with Phonological and Lexical Models", *Proc. of ICASSP 96*, pp. 3507-3510.
- [9] E. Bocchieri and J. G. Wilpon, (1992), "Discriminative Analysis for Feature Reduction in Automatic Speech Recognition," *Proc. of ICASSP 92*, Vol. I, pp. 501-504.
- [10] F. Pereira, M. Riley and R. Sproat, (1994, "Weighted Rational transduction and their application to human language processing", *DARPA Workshop on Human Language Tech*, Princeton, NJ, 1994.
- [11] D. Hindle, Preprint 1994.
- [12] R. C. Rose and D. B. Paul, (1990), "A Hidden Markov Model Based Keyword Recognition System", *Proc. of ICASSP 90*, Vol. I, pp. 129 - 132.
- [13] Sheryl Young, (1993), "Detecting Misrecognitions and Out-of-Vocabulary Words", *Proc. of ICASSP 94*, Vol. II, pp. 21-24.
- [14] Stephen Lowe, Anne Demedts, Larry Gillick, Mark Mandel, and Barbara Peskin, (1994 "Language Identification via Large Vocabulary Speaker Independent Continuous Speech Recognition," *Proc. of ARPA Human Language Technology Workshop*, 1994.
- [15] Sergio Mendoza, Larry Gillick, Yoshiko Ito, Stephen Lowe, and Michael Newman, (1996 "Automatic Language Identification Using Large Vocabulary Continuous Speech Recognition," *Proc. of ICASSP 96*, Vol. 2, pp. 785-788.
- [16] Mazim Rahim and Jay Wilpon, (1993 "Cepstral Mean Subtraction for Recognition of Telephone Speech," *AT&T Technical Memorandum* 1993.