ESC-TR-94-094

Technical Report
1010

Language Identification through Parallel Phone Recognition

C.S. Chou
M.A. Zissman

19 May 1995

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Department of Defense under Air Force Contract F19628-95-C-0002.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Department of Defense under Air Force Contract F19628-95-C-0002.

This report may be reproduced to satisfy needs of U.S. Government agencies.

The ESC Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tutungian

Gary Tutungian
Administrative Contracting Officer
Contracted Support Management

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

LANGUAGE IDENTIFICATION THROUGH PARALLEL PHONE RECOGNITION

C.S. CHOU
M.A. ZISSMAN
Group 24

TECHNICAL REPORT 1010

19 MAY 1995

Approved for public release; distribution is unlimited.

LEXINGTON MASSACHUSETTS

ABSTRACT

Language identification systems that employ acoustic likelihoods from language-dependent phoneme recognizers to perform language classification have been shown to yield high performance on clean speech. In this report, such a method was applied to language identification of telephone speech. Phoneme recognizers were developed for English, German, Japanese, Mandarin, and Spanish using hidden Markov models. Each of these processed the input speech and output a phoneme sequence in their respective languages along with a likelihood score. The language of the incoming speech was hypothesized as the language of the model having the highest likelihood. The main differences between this system and those developed in the past are that this system processed telephone speech, could identify up to five languages, and used phonetic transcriptions to train the language-specific models. The five-language, forced-choice recognition rate on 45-s utterances was 71.9%. On 10-s utterances the recognition decreased to 70.3%. In addition, it was found that adding word-specific phonemes to the training set had a negligible effect on language identification results.

ACKNOWLEDGMENTS

I would like to thank the people in Group 24 at MIT Lincoln Laboratory and the Department of Defense for their interest and support in my research and education.

Also, I would like to thank all the people who have given me support throughout my studies at MIT. Without the support from Dave, my friends, and my family during this time, I would not have made it to this point. But mostly I would like to thank my parents, who have made everything in my life possible. They have supported me through all my endeavors and have smiled at my success.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

In the past five years, language identification (LID) of speech messages has become an increasingly important part of digital speech processing systems. LID systems can be used as preprocessors in automatic language translators, in systems used by operators to identify the language of a caller, and in information centers at public airports and train stations.

Language identification performed by running several different language-dependent phoneme recognizers has been shown to be successful in experiments run on language pairs [1,2]. The primary purpose of this report was to determine the feasibility and performance of a parallel phoneme recognition LID system on telephone speech spoken in any of five languages. In addition, this report measured the effect of adding word-specific phonemes to each language's training set.

The rest of this report is organized as follows: Section 2 contains background information and presents several LID systems and their results. Section 3 explains the implementation and results from the baseline system, and Section 4 compares these results with those attained when the system trains on word-specific phonemes as well. Section 5 presents the results of using phone-based acoustic likelihoods to perform five-language identification. Finally, Section 6 summarizes the results and suggests future research directions.

2. PREVIOUS WORK

2.1 Introduction

Several language identification methods, including a phoneme recognition system similar to the one used in this report, have already been developed and tested in the past. In this section, a few of the major language identification systems are presented. Each subsection details a specific LID system, including the model, method, type of data, training data, and results. In addition, where appropriate, observations are made that pertain directly to this report.

2.2 Language-Dependent Phone Recognition

Lamel and Gauvain [1] developed an LID system based on phoneme recognition. Their system processed incoming speech in parallel through French and English phone networks. The phone models were three-state, left-to-right, continuous-density HMMs with Gaussian mixture observation densities. The language of the speech was hypothesized as the language of the phone network with the highest likelihood score. A graphic representation of this system is shown in Figure 1. Lamel and Gauvain used four corpora containing read speech to train and test their system. These were the Base de Données des Sons du Français (BDSONS) corpus and the BREF corpus for French speech, and the DARPA Wall Street Journal and TIMIT corpora for English speech. They achieved a 99% accuracy rate with 2 s of the clean speech. However, this result may not be as conclusive as it first appears as the speech used for training and testing was not collected consistently.

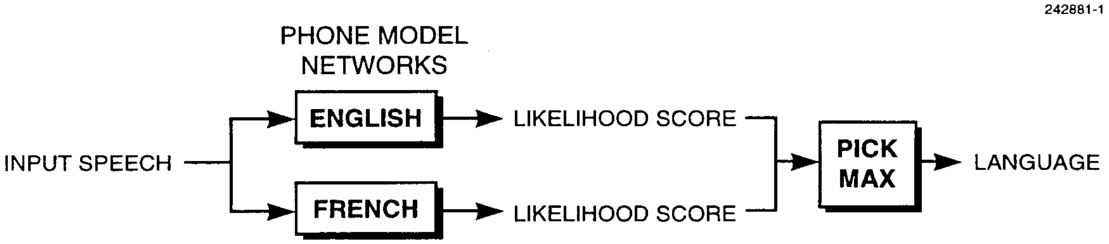

Figure 1. Lamel-Gauvain LID system.

More recently, Lamel and Gauvain performed language identification on the Oregon Graduate Institute (OGI) Multi-Language Telephone Speech (OGI-TS) Corpus [3], the same corpus used in this report; however, their language-specific models were trained without the use of phone transcriptions. Rather, they used speaker-independent, context-independent phone models, trained using the NTIMIT [4] corpus, to label the training data and then used these labels to train language-specific, OGI phone models. They achieved a 59% accuracy rate for 10-language identification on 10-s utterances. In comparison with their previous French/English efforts, two-way French/English language identification using this method and the OGI corpus operated with 82% accuracy [5].

Some of the advantages of parallel phone recognition are that it can [6]:

- Take advantage of phonotactic constraints, i.e., the restrictions found on phoneme sequences for different languages.
- Be integrated easily into existing recognizers based on phone models.

This system also has several disadvantages in that it:

- Requires phonetically or orthographically labeled training speech in all languages.
- May require a great deal of computation, i.e., a phone recognizer must be run in each language of interest.

2.3 Language-Independent Phone Recognition Followed by Language Modeling

Hazen and Zue [7] developed an automatic language identification system that incorporated separate models for the phonotactic, prosodic, and acoustic information of each language. Their system employs an English front-end phone recognizer followed by n-gram language modeling in each language to be recognized. When trained and tested using all 10 languages of the OGI-TS corpus, they initially achieved an overall system performance of 57% on 45-s utterances and 46% on 10-s utterances on the National Institute of Science and Technology (NIST) 1993 evaluation data.[1] Subsequently, they have improved performance to 69% on 45-s utterances and 64% on 10-s utterances as reported at the NIST 1994 evaluation.

A recently developed method used at MIT Lincoln Laboratory for language identification is the parallel phoneme recognition followed by language modeling (PRLM-P) method, which involves the use of multiple phoneme recognizers with n-gram language models [8]. The sequence of phonemes output from each phoneme recognizer is compared with n-gram language models computed from training speech for each of the various languages under consideration. The language with the highest likelihood score is determined to be the language of the speech. It is not necessary to have a phone recognizer in each language to be identified; rather, one language model per front-end recognizer per input language is trained, as shown in Figure 2. At the 1994 March NIST evaluation, this system exhibited the best identification performance across many different

[1] The 1993 and 1994 NIST evaluation techniques and results can be obtained from Dr. Alvin Martin at NIST in Gaithersburg, MD.

test scenarios. For example, OGI telephone speech language identification performance was 80% for 45-s test utterances and 70% for 10-s utterances. Average language pair performance was 95% for 45-s utterances and 92% for 10-s utterances.

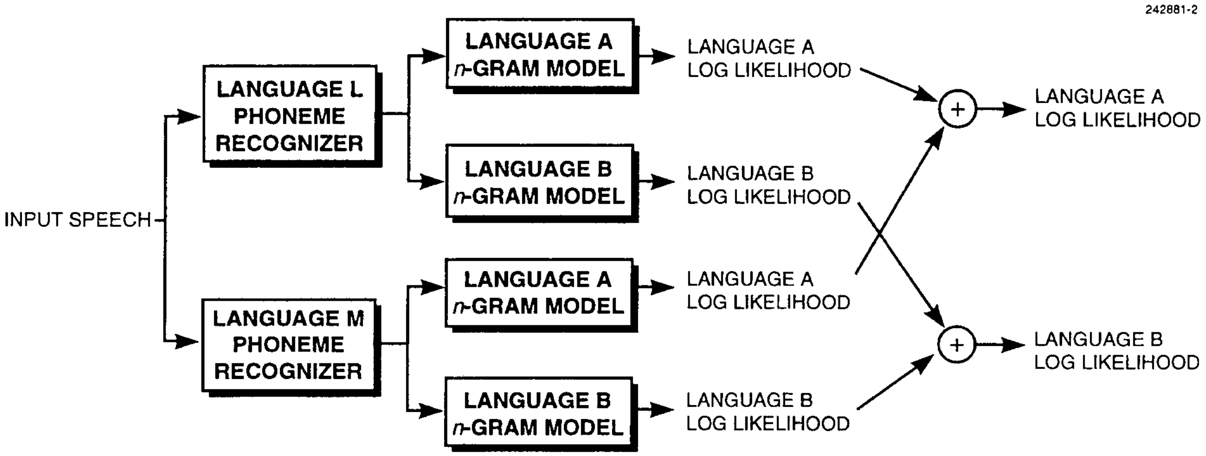
242881-2
INPUT SPEECH
LANGUAGE L PHONEME RECOGNIZER
LANGUAGE M PHONEME RECOGNIZER
LANGUAGE A n-GRAM MODEL
LANGUAGE B n-GRAM MODEL
LANGUAGE A n-GRAM MODEL
LANGUAGE B n-GRAM MODEL
LANGUAGE A LOG LIKELIHOOD
LANGUAGE B LOG LIKELIHOOD
LANGUAGE A LOG LIKELIHOOD
LANGUAGE B LOG LIKELIHOOD
+
+
LANGUAGE A LOG LIKELIHOOD
LANGUAGE B LOG LIKELIHOOD

Figure 2. MIT Lincoln Laboratory PRLM-P system.

2.4 Phonetic-Class-Based Approaches

Phonetic-class-based approaches are very similar to phoneme-based approaches. The main difference is in the types of units that are recognized in each system. In phonetic-class-based approaches, the objective is the recognition of broad phonetic class elements (i.e., vowel; fricative; stop; pre-, inter-, and post-vocalic sonorant; silence or background noise, etc.). The system requires phonetic-class-labeled data for training. The smaller number of units relative to phoneme-based approaches makes the class recognition faster and more accurate.

House and Neuburg were the first to propose the phonetic-class-based approach [9]. They developed an HMM for each language. A maximum likelihood decision rule was then used to hypothesize the language of the incoming speech. They tested their system on eight phonetic texts of the same fable, each in a different language. These fables were reduced to four-character alphabets and tested on the statistical models of each language.

Muthusamy and Cole [10] developed a similar system that segmented the speech into seven broad phonetic categories and classified the feature measurements from these categories. They trained and tested their system on the 10 languages in the OGI-TS Corpus, achieving 66% accuracy on 45-s utterances and 48% accuracy on 10-s utterances at the NIST 1993 evaluation.

2.5 Frame-Based Approaches

Frame-based approaches differ from both preceding approaches in that they do not require labeled data for training. Goodman [11] applied this approach to a very noisy, six-language database. He used a formant-cluster algorithm in which linear prediction coding (LPC)-based formants were extracted and the Euclidean distance measure was used to determine the closest clusters to the input vector. This distance was accumulated and the language was determined to be the one with the smallest total distance.

Sugiyama [12] and Nakagawa [13] performed vector quantization (VQ) classification on LPC features. Sugiyama investigated the differences between using a VQ codebook for each language and a universal VQ codebook for all languages. The algorithms had 65% and 80% recognition rates, respectively. Nakagawa investigated the use of a codebook with a continuous HMM (CHMM), a discrete HMM (DHMM), and an HMM with continuous mixture density output probability functions (CMDF). The CHMM and CMDF had comparable performance, with an 86.3% accuracy rate, while the DHMM had worse results, with a 47.6% accuracy rate.

Zissman studied the use of continuous observation, ergodic HMMs with tied Gaussian observation probability densities [14]. The HMMs were trained for each language using the mel-weighted cepstra and mel-weighted delta cepstra taken from the training speech. The same feature vectors were extracted from the test speech to test the HMMs. Likelihood scores for each language were generated from which the language of the incoming speech was determined. Ten-language classification performance on the OGI Corpus was 53% on 45-s utterances and 50% on 10-s utterances on the NIST 1993 data. Generally, the multistate HMMs performed no better than simpler Gaussian mixture classifiers.

3. BASELINE SYSTEM

3.1 Introduction

A system similar to the Lamel and Gauvain LID system was developed as a baseline for this report. Phoneme recognizers were developed for English and Spanish. The baseline system was used to determine the best implementation for performing language identification. One of the components investigated was the set of phonemes on which the system was trained. In particular, the effect of the addition of word-specific phonemes was determined. This section explains the implementation and results of the baseline system. Section 4 compares these results with those obtained when word-specific phonemes are included.

3.2 The System

The baseline system was a parallel phoneme recognition system similar to that of Lamel and Gauvain (discussed in Section 2). Incoming speech was processed in parallel through an English phone model network and a Spanish phone model network. The baseline system used the difference in likelihood scores to sort the messages according to their likelihood of being either English or Spanish. A graphic representation of the baseline system is shown in Figure 3.

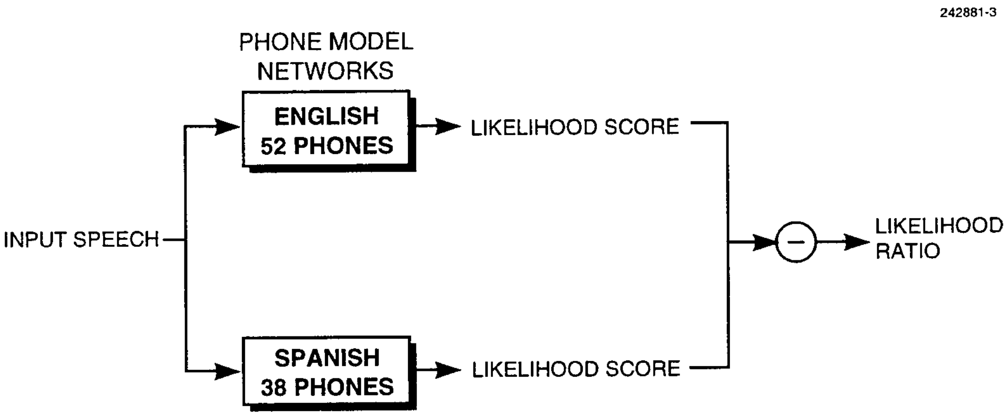

Figure 3. Baseline system.

The HMM Toolkit (HTK) [15] was used to build the phoneme recognizers. Mel-weighted cepstra and mel-weighted delta cepstra observation streams were processed statistically independently of each other. Each phone model had three emitting states, and each state used one six-component Gaussian mixture model to model the cepstra and another six-component model for the delta cepstra. Diagonal variances were employed. Training was performed using the Baum-Welsh algorithm. Recognition was performed using a Viterbi recognizer, which produced the most likely phone sequence along with that sequence's log likelihood score normalized by the number of frames.[2] The intermodel log transition probabilities between two connected phoneme models were defined as:

$$s\log[P(j|i)] \quad , \tag{1}$$

where s is the grammar scale factor, the value of which was set during preliminary tests. $P(j|i)$ was defined using bigram probabilities determined from the phone labels during training. The phone networks contained monophones and the top 100 most frequently occurring right-diphones[3] from the training data for both languages.

3.3 OGI Telephone Speech Corpus

The OGI-TS Corpus was used to train and test the system [3]. It was designed to support research on automatic language identification and multilanguage speech recognition. Each caller gave up to nine separate responses, ranging from single words and short topic-specific descriptions to 60 s of unconstrained spontaneous speech. The utterances were spoken over commercial telephone lines by speakers in English, Farsi (Persian), French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese. The speech files for each language were divided into 50 training messages, 20 development test messages, and 20 evaluation test messages.

Because the parallel phoneme recognizers used in this system required phonetically labeled data for training, only the 45-s-long "story-before-the-tone" (story-bt) utterances could be used, as these were the only labeled data in the corpus. To get the input speech into a more useful format for training, the 45-s story-bt utterances were broken down into smaller segments by removing silences and superfluous sounds. Thus the original 44 English and 48 Spanish training speech files were broken into 677 and 806 smaller files, respectively, mostly under 6 s in length. The final amounts of training as well as testing data are given in Table 1.[4] After cepstra and delta cepstra vectors were computed from input files, RelAtive SpecTrAl (RASTA) filtering [16] was used as a front-end processor to remove the effects of variable telephone line channels. In all, these data were used

[2] For the rest of this report, the term "likelihood score" refers to these normalized log likelihood scores.

[3] A right-diphone is a right context-dependent phone model.

[4] It appears that there are more testing than training data because the silences were removed from the training data and left in the testing data.

to train 52 English monophones and 38 Spanish monophones, as well as the 100 most frequently occurring diphones in each language.

TABLE 1

Amounts of English and Spanish Training and Testing Data

Language	Training Data	Testing Data
English	27.23 min	27.03 min
Spanish	26.29 min	24.53 min

Testing was carried out according to the NIST April 1993 specification. Sixty-three English and 54 Spanish 10-s utterances were used.

3.4 Performance Metrics

Rather than assessing the system by performing language identification between the two languages, the likelihood ratio output from the baseline system was used to generate receiver operating curves (ROCs) and their figures of merit (FOMs). This method was preferable because likelihood score biases had been observed in previous tests of such systems at Lincoln. By taking the difference in the likelihood scores, this bias problem was eliminated.

ROCs were generated by plotting the probability of detection, P_D, on the y-axis versus the probability of false alarm, P_F, on the x-axis for all possible score thresholds. The area under this curve is the FOM. For an ideal system, $P_D = 1$ and $P_F = 0$, so the ROC would be two straight lines from (0,0) to (0,1) to (1,1) and the FOM would be equal to one. The closer a system's ROC is to this ideal curve (i.e., the closer the FOM is to one), the better the system performance.

3.5 Results

The grammar scale factor, s, was set after running some preliminary tests to determine its effect on language identification. Several different tests were run with the only difference being this factor. The value of this factor in the various tests along with the FOM from the resulting ROCs are given in Table 2. The ROCs for these tests are shown in Figure 4 for the case of English targets and Spanish background. These data show that performance was relatively insensitive to s, so $s = 3$ was used in all subsequent tests. With $s = 3$, the baseline system had a 0.979 FOM.

TABLE 2

Grammar Scale Factor Values and LID Figures of Merit

Grammar Scale Factor	Figure of Merit
s = 1	0.976
s = 3	0.979
s = 5	0.978
s = 10	0.966

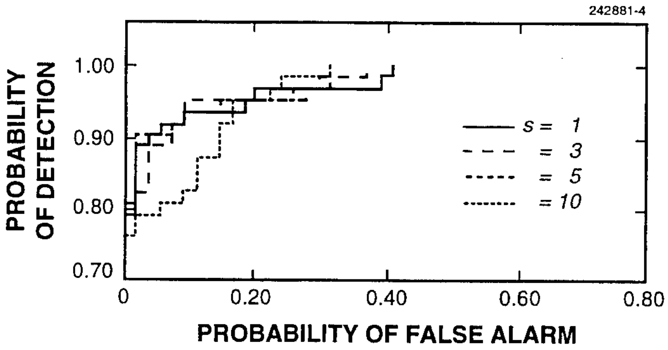

Figure 4. ROCs for various grammar scale factor values; 10-s utterances; target = English, background = Spanish.

4. WORD-SPECIFIC PHONEME TESTS

4.1 Introduction

The inclusion of word-specific phone models was investigated to determine whether it would improve the performance of the baseline system. These new phone models were trained only on occurrences in certain words. For example, the word *the* is usually composed of two phones, /DH/ and /AX/. Considering the /DH/ phone, a general /DH/ phone was trained on occurrences of /DH/ in all words other than *the*, such as *this* and *there*. A separate phone, /DH-the/ was trained from occurrences of *the*. Word-specific phone models of commonly occurring words were incorporated into the baseline system to see how they affected the system's language identification performance.

To incorporate this change into the baseline system, the commonly occurring words needed to be manually tagged in the segmented input data. The top five most frequently occurring words in spoken English are [17]

- I
- and
- the
- to
- that.

The top six[5] most frequently occurring words in Spanish are [18]

- de
- el
- la
- y
- a
- en.

Because the OGI training speech is phonetically, but not orthographically, transcribed, listeners tagged occurrences of these frequent words manually and then recorded the OGI phonetic labels corresponding to each word occurrence. For the word *the*, several different phonetic expansions might be observed, e.g., /DH AX/, /DH IY/, etc. The word-specific phonemes, along with the

[5] The sixth word, *en*, was added to the list after initial tagging of the training data had begun and it was found to occur as often as the other words in the list.

number of occurrences of each, are given in Tables 3 and 4. These tables also show the percentage of all phones that were included in these words. With the addition of these word-specific phonemes, the original monophone list was expanded from 52 to 76 monophones for English and from 38 to 52 monophones for Spanish.

4.2 Results

Running on English versus Spanish data as described in Section 3.3, this word-specific phone system also had a 0.979 FOM. The ROC for this system is compared with that of the baseline system in Figure 5.

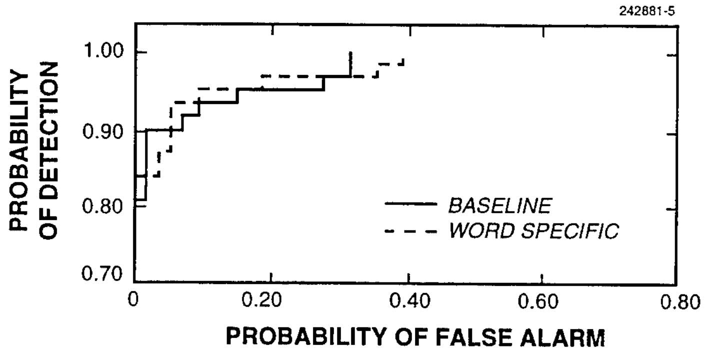

Figure 5. ROC for word-specific phoneme system versus baseline system; 10-s utterances; target = English, background = Spanish.

The inclusion of the word-specific phonemes brought no improvement in language identification, perhaps because the word-specific phones covered only approximately 5% of the data. However, to measure the small-scale effectiveness of this change, further analysis was done. In particular, the number of times the system correctly or incorrectly detected the word-specific phonemes was determined. This result was compared with the phonemes specified by the baseline system. The results of this analysis are given in Tables 5 and 6.

These results indicate that the baseline system actually recognized the word-specific phonemes better than the system that was trained on them. In particular, almost all the word-specific phonemes in both English and Spanish were recognized by both systems or by neither system. Of

TABLE 3

Phonetic Breakdown and Frequency of Occurrences of English Word-Specific Phones

Word	Phonetic Transcription	Frequency in Training Data	Frequency in Testing Data	Percentage of All Phones in Training Data	Percentage of All Phones in Testing Data
I	/AY-I/	77	57	0.3197%	0.8463%
	/AE-I/	3	3		
	/AH-I/	4	10		
and	/AE-and/	83	39	0.7687%	1.5720%
	/EH-and/	2	11		
	/N-and/	82	50		
	/VCL-and/	14	13		
	/D-and/	21	17		
the	/DH-the/	247	69	1.8610%	1.6560%
	/TH-the/	4	2		
	/IH-the/	22	14		
	/AX-the/	104	15		
	/AH-the/	48	24		
	/IY-the/	64	13		
to	/T-to/	122	36	0.8410%	0.7496%
	/AH-to/	10	1		
	/AX-to/	22	7		
	/IX-to/	16	2		
	/UW-to/	51	16		
that	/DH-that/	64	24	0.5251%	0.5441%
	/AH-that/	6	3		
	/AE-that/	42	14		
	/CL-that/	15	2		
	/T-that/	11	2		
Total		1134	444	4.3155%	5.3680%

TABLE 4

Phonetic Breakdown and Frequency of Occurrences of Spanish Word-Specific Phones

Word	Phonetic Transcription	Frequency in Training Data	Frequency in Testing Data	Percentage of All Phones in Training Data	Percentage of All Phones in Testing Data
de	/D-de/	77	15	1.3790%	1.2650%
	/DX-de/	88	32		
	/EY-de/	161	48		
el	/EY-el/	62	27	0.5924%	0.8124%
	/L-el/	78	34		
la	/L-la/	111	44	0.9351%	1.1850%
	/AA-la/	110	45		
y	/EY-y/	17	8	0.6135%	0.7591%
	/IY-y/	126	48		
	/Y-y/	2	1		
a	/AA-a/	47	9	0.1989%	0.1199%
en	/EY-en/	98	51	0.8674%	1.3050%
	/N-en/	86	41		
	/NG-en/	21	6		
Total		1084	409	4.5863%	5.4464%

the word-specific phonemes that were only recognized by one, the baseline system detected more than the word-specific phoneme system.

4.3 Conclusion and Future Work

Preliminary experiments were run to determine the effect of adding word-specific phonemes to the training set. The evidence seems to weigh in favor of leaving out the word-specific phonemes, especially considering the additional man-hours needed to tag them. If there were larger orthographically transcribed databases, a word-spotting or word-recognition approach to language identification could be pursued. Investigating this approach with the current OGI database, which may be too small to train word-specific phone models and is not orthographically transcribed, would be difficult.

TABLE 5

Comparison of Recognition Performance for English

Figure of Merit

Basis	Baseline System	Word-Specific Phone System
Overall	0.979	0.979

Phone Recognition Performance on "Keywords"

Basis	Baseline System	Word-Specific Phone System[a]
Overall	46.8%	40.8%
Recognized by this system only	10.6%	4.62%
Recognized by neither system	48.6%	

[a] Includes recognizing the base phone only, i.e., if the word-specific phone system recognized /AE/ when the actual word was /AE-and/, it was counted as correctly recognizing the phone.

TABLE 6

Comparison of Recognition Performance for Spanish

Figure of Merit

Basis	Baseline System	Word-Specific Phone System
Overall	0.979	0.979

Phone Recognition Performance on "Keywords"

Basis	Baseline System	Word-Specific Phone System[a]
Overall	60.9%	59.1%
Recognized by this system only	7.40%	0.77%
Recognized by neither system	33.4%	

[a] Includes recognizing the base phone only, i.e., if the word-specific phone system recognized /EY/ when the actual word was /EY-en/, it was counted as correctly recognizing the phone.

5. FURTHER EXPERIMENTS USING ACOUSTIC LIKELIHOODS

5.1 Introduction

This section details the development of the complete LID system using phone-based acoustic likelihoods. Phoneme recognizers were developed in English, German, Japanese, Mandarin, and Spanish and were used to create an LID system similar to that of the baseline. The system was built and tested to determine the feasibility and performance of a parallel phoneme recognition system on telephone speech.

5.2 The System

The LID system developed for these tests was a parallel phoneme recognition (PPR-C)[6] similar to that of the baseline system described in Section 3. Incoming speech was processed in parallel through English, German, Japanese, Mandarin, and Spanish phone model networks. The language of the incoming speech was hypothesized as the language of the model having the highest likelihood. A graphic representation of this system is shown in Figure 6.

As was done for the baseline system, the 45-s story-bt training utterances in German, Japanese, and Mandarin were broken down into smaller segments and the superfluous sounds were removed. The final amounts of training and testing data for all five languages are given in Table 7[7] along with the number of monophones trained in each language. The implementation of this system is the same as that of the baseline system that was detailed in Section 3. However, each of the five phone networks used when testing this system contained only monophones.

5.3 Performance Measures

Five-way language classification was used to assess the performance of the system. The likelihood scores output from the system were adjusted before language identification was performed to address the bias issue that had been noticed in previous language identification tests. An adjustment was made by postprocessing the raw likelihood scores such that for each recognizer, the mean of the scores from all messages processed by the recognizer was set to zero. Thus the adjustment took the form of a recognizer-dependent addition or subtraction. The resulting likelihood scores were compared and the language of the model with the highest likelihood score was hypothesized as the language of the incoming speech. Language identification performance is given by the ratio of the number of speech files the language of which was correctly identified divided by the total number of files.

[6] Parallel phoneme recognition performed by Chou.

[7] It appears that there are more testing than training data because the silences were removed from the training data and left in the testing data.

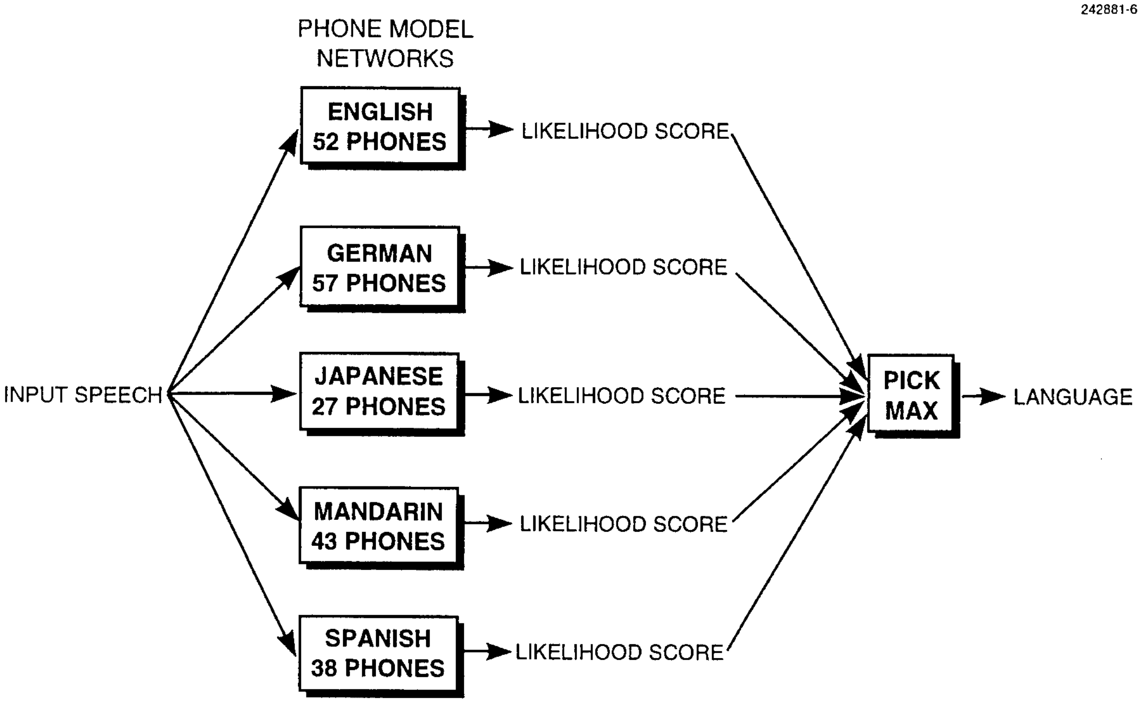

Figure 6. LID system using acoustic likelihoods.

TABLE 7

Amounts of Training and Testing Data for Five-Language Identification System

Language	Training Data	Testing Data	No. of Monophones
English	27.23 min	27.03 min	52
German	24.45 min	26.54 min	57
Japanese	23.44 min	25.16 min	27
Mandarin	17.69 min	26.93 min	43
Spanish	26.29 min	24.53 min	38

5.4 Results

Running according to the NIST 1993 specifications, the PPR-C system attained a five-language recognition rate of 70.3% correct on the 10-s utterances. On 45-s utterances, this recognition rate increased to 71.9%. Table 8 shows the five-language confusion matrix. Table 9 compares these results with those of Zissman's PRLM-P system, which was described briefly in Section 2. When the PRLM-P system was tested on the same five languages, it achieved a language recognition rate of 75.7% on the 10-s utterances and 86.5% on the 45-s utterances. The standard deviations (σ), estimated according to a Bernoulli model, are shown in the bottom row of Table 9.

TABLE 8

Five-Language Confusion Matrices

10-s Utterances Test					
	Hypothesized Language				
Actual Language	**English**	**German**	**Japanese**	**Mandarin**	**Spanish**
English	47	10	3	1	2
German	12	46	2	0	3
Japanese	1	0	53	1	2
Mandarin	7	14	8	26	4
Spanish	3	6	9	0	36
45-s Utterances Test					
	Hypothesized Language				
Actual Language	**English**	**German**	**Japanese**	**Mandarin**	**Spanish**
English	12	6	0	0	0
German	1	16	0	0	1
Japanese	0	0	16	0	1
Mandarin	2	6	1	9	1
Spanish	0	5	1	0	11

Additional analysis was done comparing the two systems' two-language identification results averaged over the 10 language pairs. These results are also given in Table 9. Again, it is evident that the Chou PPR-C system developed in this report has lower accuracy than the Zissman PRLM-P approach.

TABLE 9

Identification Results

System	Five Language		Two Language	
	45 s	10 s	45 s	10 s
PRLM-P	86.5%	75.7%	94.7%	89.2%
PPR-C	71.9%	70.3%	88.0%	86.5%
σ	5%	3%	2%	1%

English/Japanese/Spanish experiments were also performed on the Chou PPR-C system for further comparison with Zissman's PRLM-P and PPR systems. These results are presented in Tables 10 and 11 and show that Chou's PPR-C system has comparable performance with Zissman's PPR and PRLM-P systems on each of the three language pairs. This outcome is expected because the two systems are trained and tested on the same data and are using basically the same approach.

TABLE 10

English/Japanese/Spanish Language Pair Identification Results

System	Two-Language Identification							
	English/Spanish		English/Japanese		Japanese/Spanish		Average	
	45 s	10 s	45 s	10 s	45 s	10 s	45 s	10 s
PRLM-P	97.1%	88.0%	91.4%	90.0%	94.1%	90.1%	94.2%	89.4%
PPR	97.1%	92.3%	94.3%	92.5%	85.3%	87.4%	92.2%	90.7%
PPR-C	97.1%	91.5%	94.3%	90.8%	85.3%	86.5%	92.2%	89.3%
σ							3%	2%

The results of three-language (English/Japanese/Spanish) identification are given in Table 11. The Zissman PRLM-P system had the best results (with the PPR system performing slightly below), and the Chou PPR-C system had the worst results (slightly below the Zissman PPR system). Although the statistical significance of the difference is marginal, the discrepancy between the two PPR systems could be attributed to Zissman's PPR system using the monophones plus the top 100 most commonly occurring diphones from the training data, whereas Chou's PPR-C system used only monophones.

TABLE 11

English/Japanese/Spanish Three-Language Identification Results

System	45 s	10 s
PRLM-P	92.3%	85.1%
PPR	86.5%	85.1%
PPR-C	82.7%	82.2%
σ	6%	3%

5.5 Conclusion

The results from the English/Japanese/Spanish experiments validate the Chou PPR-C system because these results are comparable with those of Zissman's PPR tests. In addition, both PPR systems had comparable results with Zissman's PRLM-P system. Thus for identifying up to three languages, the method of using phone-based acoustic likelihoods is good and produces relatively accurate results.

The results for the five-language tests show larger differences in the performance between Chou's PPR-C and Zissman's PRLM-P systems. This discrepancy seems to indicate that as the number of languages increases, the PPR-C system may have inferior recognition capabilities. Because there is some evidence that adding context-dependent diphones can improve PPR performance, future comparisons should be performed using context-dependent phone models in PPR systems.

6. CONCLUSION

This work demonstrates that language identification on telephone speech using phone-based acoustic likelihoods is feasible but does not yet produce results comparablewith other systems. On three-language identification, Chou's PPR-C system developed here had similar results to Zissman's PRLM-P and PPR systems. However, for five-language identification the PPR-C system attained a recognition rate of 71.9% correct, much lower than the 86.5% correct achieved by the PRLM-P system. Adding context-dependent phones to the phone recognizers might improve PPR performance and should be the subject of future work. Additionally, it was shown that simple addition of commonly occurring word-specific phonemes did not improve PPR performance. Perhaps with the advent of larger multilanguage speech corpora, word-specific modeling approaches will be more appropriate.

REFERENCES

1. L.F. Lamel and J-L. Gauvain, "Cross-lingual experiments with phone recognition," *ICASSP '93 Proc.* **2**, 507–510 (April 1993).

2. Y. Muthusamy, K. Berkling, T. Arai, R. Cole, and E. Barnard, "A comparison of approaches to automatic language identification using telephone speech," *Proc. Eurospeech 93* **2**, 1307–1310 (September 1993).

3. Y.K. Muthusamy, R.A. Cole, and B.T. Oshika, "The OGI multi-language telephone speech corpus," *ICSLP '92 Proc.* **2**, 895–898 (October 1992).

4. C.R. Jankowski et al., "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," *ICASSP '90 Proc.* (April 1990).

5. L.F. Lamel and J-L. Gauvain, "Identifying non-linguistic speech features," *Proc. Eurospeech 93* **1**, 23–30 (September 1993).

6. J-L. Gauvain and L.F. Lamel, "Identification of non-linguistic speech features," *DARPA Human Lang. Technol. '93 Proc.* (March 1993).

7. T.J. Hazen and V.W. Zue, "Automatic language identification using a segment-based approach," *Proc. Eurospeech 93* **2**, 1303–1306 (September 1993).

8. M.A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and *n*-gram modeling," *ICASSP '94 Proc.*, 305–308 (April 1994).

9. A.S. House and E.P. Neuburg, "Toward automatic identification of the language of an utterance, I, Preliminary methodological considerations, *J. Acoust. Soc. Amer.*, 708–713 (September 1977).

10. Y.K. Muthusamy and R.A. Cole, "Automatic segmentation and identification of ten languages using telephone speech," *ICSLP '92 Proc.* **2**, 1007–1010 (October 1992).

11. F.J. Goodman, A.F. Martin, and R.E. Wohlford, "Improved automatic language identification in noisy speech," *ICASSP '89 Proc.* **1**, 528–531 (May 1989).

12. M. Sugiyama, "Automatic language recognition using acoustic features," *ICASSP '91 Proc.* **2**, 813–816 (May 1991).

13. S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text-independent language identification by HMM," *ICSLP '92 Proc.* **2**, 1011–1014 (October 1992).

14. M.A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," *ICASSP '93 Proc.* **2**, 399–402 (April 1993).

15. P.C. Woodland and S.J. Young, "The HTK tied-state continuous speech recogniser," *Proc. Eurospeech 93* **3**, 2207–2210 (September 1993).

REFERENCES
(Continued)

16. H. Hermansky et al., “RASTA-PLP speech analysis technique,” *ICASSP ’92 Proc.* **1**, 121–124 (March 1992).

17. H. Dahl, *Word Frequencies of Spoken American English*, Essex, Conn.: Verbatim (1979).

18. A. Juilland and E. Chang-Rodriguez, *Frequency Dictionary of Spanish Words*, The Hague: Mouton & Co. (1964).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)

2. REPORT DATE
19 May 1995

3. REPORT TYPE AND DATES COVERED
Technical Report

4. TITLE AND SUBTITLE
Language Identification through Parallel Phone Recognition

5. FUNDING NUMBERS
C — F19628-95-C-0002
PR — 279-2-202

6. AUTHOR(S)
Christine S. Chou and Marc A. Zissman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lincoln Laboratory, MIT
244 Wood Street
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER
TR-1010

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Department of Defense
Washington, DC 20301-7100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ESC-TR-94-094

11. SUPPLEMENTARY NOTES
None

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

Language identification systems that employ acoustic likelihoods from language-dependent phoneme recognizers to perform language classification have been shown to yield high performance on clean speech. In this report, such a method was applied to language identification of telephone speech. Phoneme recognizers were developed for English, German, Japanese, Mandarin, and Spanish using hidden Markov models. Each of these processed the input speech and output a phoneme sequence in their respective languages along with a likelihood score. The language of the incoming speech was hypothesized as the language of the model having the highest likelihood. The main differences between this system and those developed in the past are that this system processed telephone speech, could identify up to five languages, and used phonetic transcriptions to train the language-specific models. The five-language, forced-choice recognition rate on 45-s utterances was 71.9%. On 10-s utterances the recognition decreased to 70.3%. In addition, it was found that adding word-specific phonemes to the training set had a negligible effect on language identification results.

14. SUBJECT TERMS
language identification
phone recognition
acoustic likelihoods
telephone speech
hidden Markov models

15. NUMBER OF PAGES
40

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Same as Report