Language Identification through Parallel Phone Recognition

by

Christine S. Chou

Submitted to the Department of
Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements
for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

May 11, 1994

Signature of Author _____
Department of Electrical Engineering and Computer Science
May 11, 1994

Certified by _____
Dr. Marc A. Zissman
Lincoln Laboratory
Thesis Supervisor

Accepted by _____
[redacted], Department Committee on Undergraduate Theses
F. R. Morgenthaler

Language Identification through Parallel Phone Recognition

by

Christine S. Chou

Submitted to the Department of Electrical Engineering and Computer Science on May 11, 1994 in partial fulfillment of the requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

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 thesis, 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 second utterances was 71.9%. On ten second 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.

Thesis Supervisor: Dr. Marc A. Zissman
Title: Staff, MIT Lincoln Laboratory

Acknowledgements

Many thanks are due to my thesis advisor, Dr. Marc Zissman, who spent many hours going over the details of my thesis with me. Without his help and knowledge, I would not have been able to complete this thesis.

I would also 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.

Finally, I would like to thank all of 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 of my endeavors and have only smiled at my success.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the different nations of the world begin to interact more frequently, language identification of speech messages is becoming an increasingly important part of digital speech processing systems. Language identification 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 on clean speech [9, 10]. From the likelihood scores output from each system, the language of the speech can be determined. This was the language identification method implemented in this thesis. Phoneme recognizers were developed for each of five languages using hidden Markov models. Each of these processed the input speech and output the most likely phoneme sequence along with a likelihood score. The language of the incoming speech was hypothesized as the language of the model having the highest likelihood. The purpose of the thesis was to determine the feasibility and performance of a parallel phoneme recognition system on telephone speech. In addition, this thesis measured the effect of adding word specific phonemes to each language's training set.

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

Chapter 2

Previous Work

2.1 Introduction

Several language identification methods, including a phoneme recognition system similar to the one used in this thesis, 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 language identification system, including the model, method, type of data, training data, and results. In addition, where appropriate, observations are made which pertain directly to this thesis.

2.2 Language Dependent Phone Recognition

Lamel and Gauvain [9] developed a language identification system based on phoneme recognition. Their system processed incoming speech in parallel through a French phone model network and an English phone model network. The phone models were three-state, left-to-right, continuous density hidden Markov models with Gaussian mixture observation densities. The language of the speech was determined to be the language represented by the phone model network with the highest likelihood score. A graphic representation of this system is shown in Figure 2.1. Lamel and Gauvain used four corpora containing read speech to train and test their system. These were the Base de Donnees des Sons du Francais (BDSONS) corpus and the BREF Corpus for French speech, and the DARPA Wall Street Journal and TIMIT Corpora for English speech. They were able to achieve an accuracy rate of 99% with two seconds 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 in a consistent same manner.

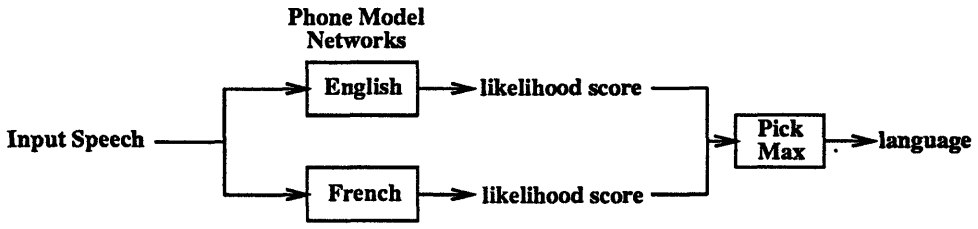

Figure 2.1: The Lamel-Gauvain Language Identification System

More recently, Lamel and Gauvain performed language identification on the OGI Multi-Language Telephone Speech Corpus [12], the same corpus used in this thesis. Their language specific models were trained without the use of phone transcriptions, however. They used speaker-independent, context-independent phone models to label the training data and then used Maximum Likelihood Estimation to estimate the language specific models. They achieved a 59% accuracy rate for 10-language identification on ten second utterances. In comparison to their previous French/English efforts, two-way French/English language identification using this method attained an accuracy rate of 82%. [4]

Some of the advantages of parallel phone recognition are given below [3]:

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

Of course, this system also has several disadvantages. The main disadvantage to this system is that it requires phonetically labeled training speech in all languages. In addition, this type of system can require a great deal of computation.

2.3 Language Independent Phone Recognition followed by Language Modeling

Hazen and Zue [6] developed an automatic language identification system which 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 ten languages of the OGI Multi-Language Telephone Speech Corpus, they initially achieved an overall system performance of 57% on 45 second utterances and 46% on ten second utterances on the NIST 1993 evaluation data[1]. Subsequently, they have improved performance to 69% on 45 second utterances and 64% on ten second 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 [16]. The sequence of phonemes output from each phoneme recognizer is compared to 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 2.2. At the 1994 March NIST evaluation, this system exhibited the best identification performance across many different test scenarios. For example, OGI telephone speech LID performance was 80% for 45 second test utterances and 70% for ten second utterances. Average language pair performance was 95% for 45 second utterances and 92% for ten second utterances.

2.4 Phonetic-Class-based Approaches

Phonetic-class-based approaches are very similar to phoneme-based approaches. The main difference is in the type 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-vocalic sonorant, inter-vocalic sonorant, post-vocalic sonorant, silence or background noise, etc.). The system requires phonetic

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

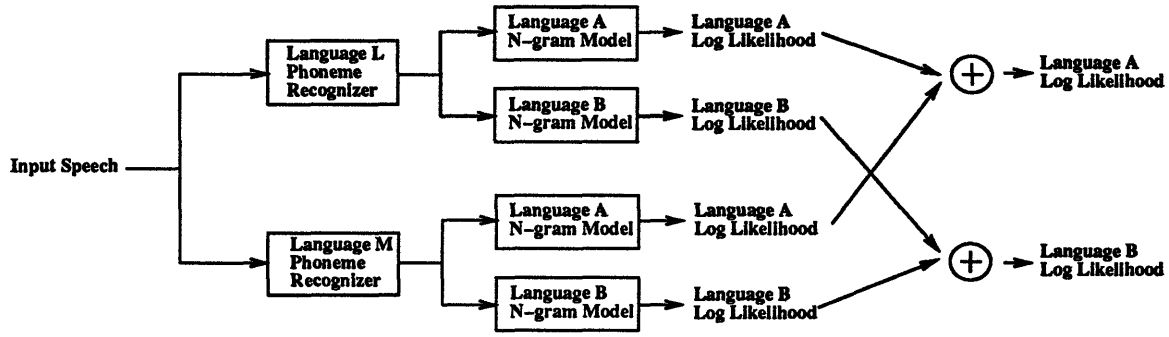

Figure 2.2: The MIT Lincoln Laboratory PRLM-P System

class labeled data for training. The smaller number of units relative to phoneme-based approaches makes the recognition faster and more accurate.

House and Neuburg were the first to propose the phonetic-class-based approach [7]. They developed a hidden Markov model 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 [11] developed a similar system which segmented the speech into seven broad phonetic categories and classified the feature measurements from these categories. They trained and tested their system on the ten languages in the OGI Multi-Language Telephone Speech Corpus, achieving 66% accuracy on 45 second utterances and 48% accuracy on ten second utterances at the NIST 1993 evaluation.

2.5 Frame-based Approaches

Frame-based approaches differ from both of the preceding approaches in that they do not require labeled data for training. Goodman [5] applied this approach to a very noisy, six language database. He used a formant-cluster algorithm in which 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 [14] and Nakagawa [13] performed vector quantization classification on LPC features. Sugiyama investigated the differences between using one VQ codebook for each language and using a universal VQ codebook for all languages. The algorithms had a 65% and 80% recognition rate respectively. Nakagawa performed a different comparison. He 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 used continuous observation, ergodic hidden Markov models with tied Gaussian observation probability densities in applying this approach [15]. The HMMs were trained for each language using the mel-weighted cepstra and 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 second utterances and 50% on ten second utterances on the NIST 1993 data. Generally, the multi-state HMMs performed no better than simpler Gaussian mixture classifiers.

Chapter 3

Baseline System

3.1 Introduction

A system similar to the Lamel and Gauvain language identification system was developed as a baseline for this thesis. 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 chapter explains the implementation and results of the baseline system. The next chapter compares these results to those obtained when word specific phonemes are included.

3.2 The System

The baseline system was a parallel phoneme recognition system similar to the Lamel and Gauvain system discussed in Chapter 2. Incoming speech was processed in parallel through an English phone model network and a Spanish phone model network. One of the main differences between this system and the Lamel-Gauvain system is that rather than picking the maximum likelihood score to determine the language of the speech, the baseline system built here took the difference in the likelihood scores and used this 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 3.1.

The Hidden Markov Model Toolkit (HTK), a toolkit for building continuous density hidden Markov model based recognizers from Cambridge University and Entropic Research Laboratory was used to build the phoneme recognizers. Mel weighted cepstra and delta cepstra observation streams were processed statistically independently

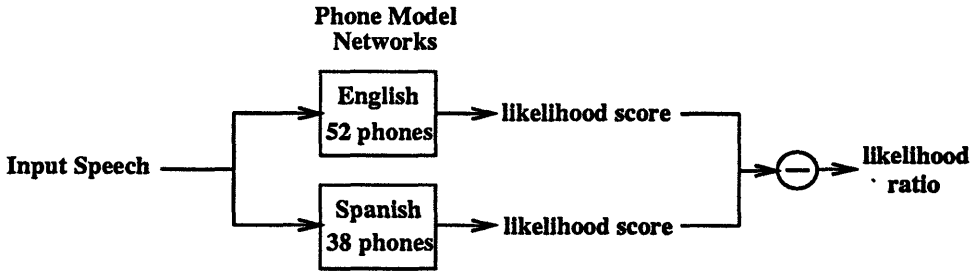

Figure 3.1: The Baseline System

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[1]. The inter-model log transition probabilities between two connected phoneme models was defined as:

$$s\log[P(j|i)] \tag{3.1}$$

where s is the grammar scale factor, whose value 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[2] from the training data for both languages.

3.3 OGI Telephone Speech Corpus

The Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus was used to train and test the system [12]. It was designed to support research on automatic language identification and multi-language speech recognition. Each caller gave up to nine separate responses, ranging from single words and short topic-specific

[1]For the rest of this thesis, the term "likelihood score" will refer to what are really these normalized log likelihood scores.

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

Table 3.1: Amounts of English and Spanish Training and Testing Data

Language	Training Data	Testing Data[a]
English	27.23 minutes	27.03 minutes
Spanish	26.29 minutes	24.53 minutes

[a]Although it appears that there is more testing data than training data, this is due to the fact that the silences were removed from the training data whereas they remained in the testing data.

descriptions to 60 seconds 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.

Since the parallel phoneme recognizers used in this system required phonetically labeled data for training, only the 45 second long "story-before-the-tone" (story-bt) utterances could be used as these were the only labeled data in the corpus. In order to get the input speech into a more useful format for training, the 45 second story-bt utterances were broken down into smaller segments by removing silences and other superfluous sounds. Thus the original 44 English and 48 Spanish training speech files were broken into 677 and 806 smaller files respectively, mostly under six seconds in length. The final amounts of training as well as testing data are given in Table 3.1. After cepstra and delta cepstra vectors were computed from input files, RASTA [2] was used as a front-end processor to remove the effects of variable telephone line channels. In all, these data were used to train 52 English monophones and 38 Spanish monophones as well as the 100 most frequently occurring diphones in each language.

Testing was carried out according to the NIST April 1993 specification. "45-sec" utterance testing refers to language identification on the 45 second story-bt utterances spoken by the development test speakers while "ten-sec" utterance testing is on ten second cuts from the "45-sec" utterances.

Table 3.2: 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

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 (ROC) and their Figures of Merit (FOM). This was preferable since 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.

Receiver operating curves 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. The area under this curve is the figure of merit (FOM). For an ideal system, $P_D = 1$ and $P_F = 0$ so the ROC curve 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 curve 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 figure of merit from the resulting receiver operating curves are given in Table 3.2. The receiver operating curves for these tests are shown in Figure 3.2. 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 figure of merit.

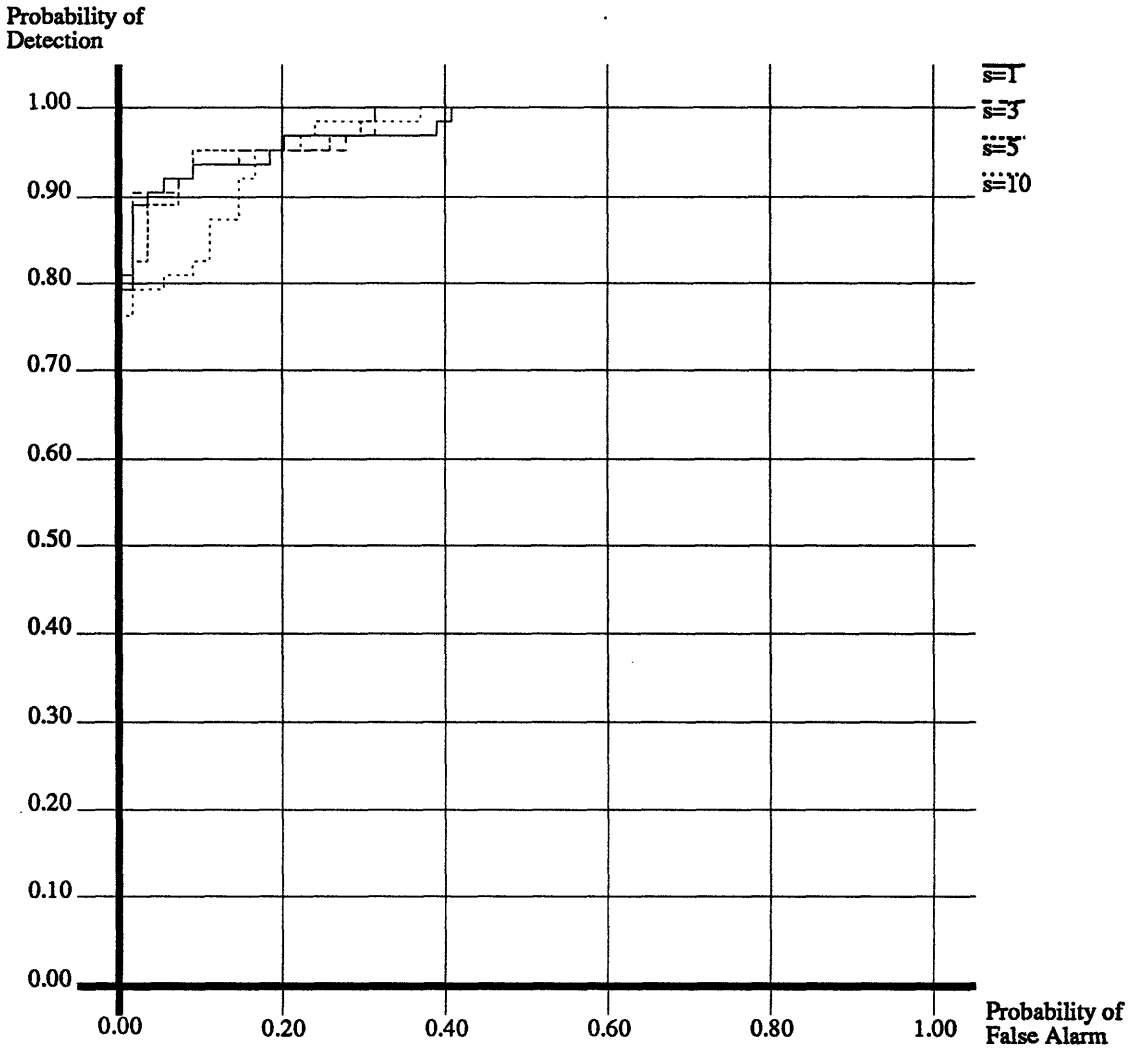

Figure 3.2: Receiver Operating Curves for various Grammar Scale Factor Values.

Chapter 4

Word Specific Phoneme Tests

4.1 Introduction

The inclusion of word specific phone models was investigated in order to determine whether this 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.

In order 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 English are: [1]

- I
- and
- the
- to
- that

The top six[1] most frequently occurring words in Spanish are: [8]

[1]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.

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

The word specific phonemes along with the number of occurrences of each are given in Table 4.1 and 4.2. These tables also show the percentage of all phones which 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 vs. Spanish data as described in Section 3.3, this word specific phone system also had a 0.979 figure of merit. The receiver operating curve for this system is compared to that of the baseline system in Figure 4.1.

The inclusion of the word specific phonemes brought only a small improvement in language identification perhaps because the word specific phones covered only around 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 was compared to the phonemes which the baseline system specified. The results of this analysis are given in Tables 4.3 and 4.4.

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

Table 4.1: 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 4.2: 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%

Table 4.3: 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]This 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.

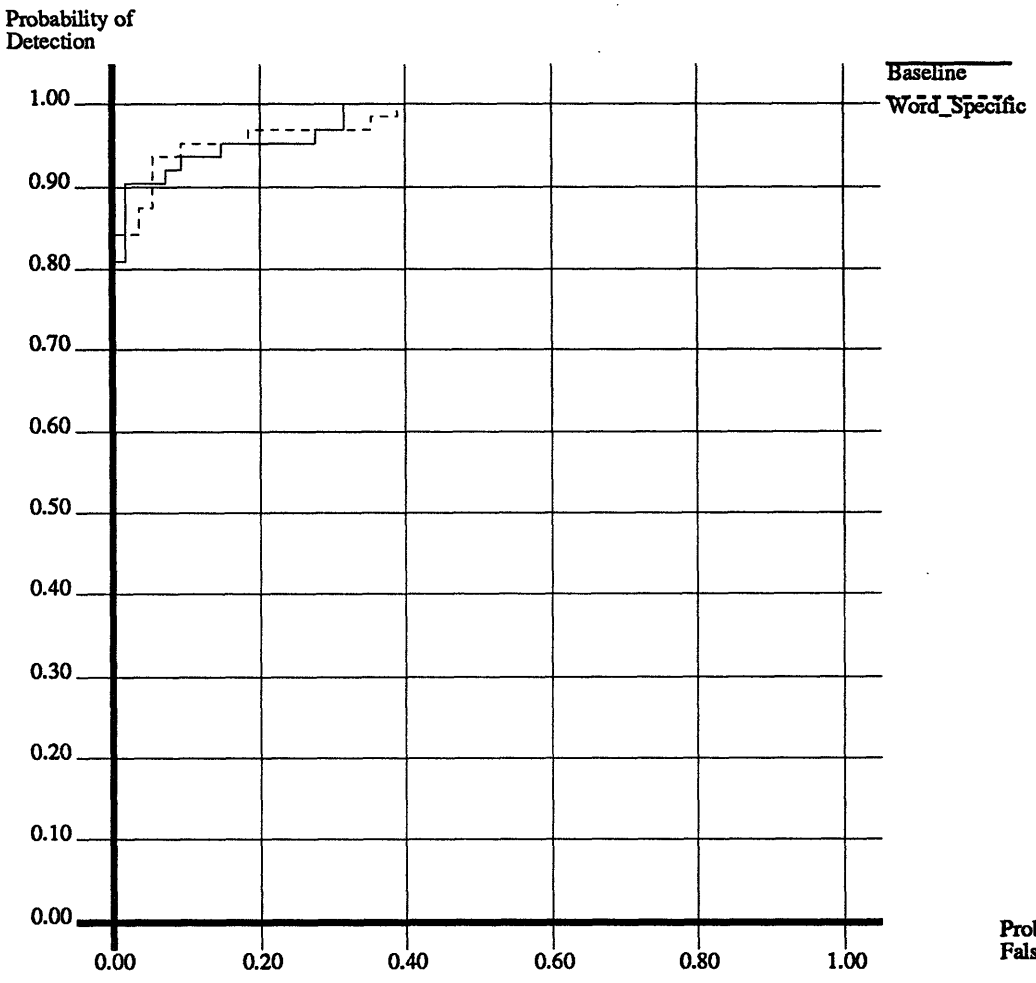

Figure 4.1: Receiver Operating Curve for Word Specific Phoneme System vs. Baseline System.

Table 4.4: 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]This 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.

only recognized by one of the systems, the baseline system detected more than the word specific phoneme system.

4.3 Conclusion and Future Work

Some 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 the word specific phonemes. If we had larger orthographically transcribed databases, a word spotting or word recognition approach to LID could be pursued. Pursuing 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.

Chapter 5

Further Experiments Using Acoustic Likelihoods

5.1 Introduction

This chapter details the development of the complete language identification system using phone-based acoustic likelihoods (PPR-C)[1]. Phoneme recognizers were developed in English, German, Japanese, Mandarin, and Spanish. These were used to create a language identification system similar to the baseline system. 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 language identification system developed for these tests was a parallel phoneme recognition system similar to the baseline system described in Chapter 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 5.1.

As was done for the baseline system, the 45 second 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 5.1 along with the number of monophones trained in each language. The implementation of this system is the same as that of

[1]PPR-C stands for Parallel Phoneme Recognition performed by Chou.

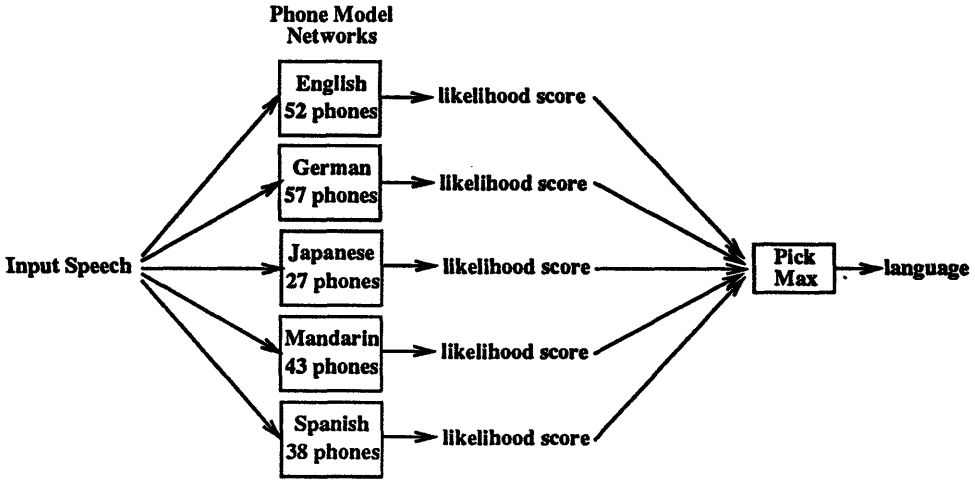

Figure 5.1: The Langugae Identification System using Acoustic Likelihoods

Table 5.1: Amounts of Training and Testing Data for Five Language Identification System

Language	Training Data	Testing Data[a]	Number of Monophones
English	27.23 minutes	27.03 minutes	52
German	24.45 minutes	26.54 minutes	57
Japanese	23.44 minutes	25.16 minutes	27
Mandarin	17.69 minutes	26.93 minutes	43
Spanish	26.29 minutes	24.53 minutes	38

[a]Although it appears that there is more testing data than training data, this is due to the fact that the silences were removed from the training data whereas they remained in the testing data.

the baseline system which was detailed in Chapter 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, however, in order to address the bias issue which had been

noticed in previous language identification tests. This was done by post processing 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 whose language was correctly identified divided by the total number of files.

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 ten second utterances. On the 45 second utterances, this recognition rate increased to 71.9%. Table 5.2 shows the five language confusion matrix. Table 5.3 compares these results to the results of Zissman's PRLM-P system which was described briefly in Chapter 2. When his PRLM-P system was tested on the same five languages, it achieved a language recognition rate of 75.7% on the ten second utterances and 86.5% on the 45 second utterances.

Additional analysis was done comparing the two system's two language identification results averaged over the ten language pairs. These are also given in Table 5.3. Once again, we see that the PPR-C system developed in this thesis has slightly lower recognition scores than Zissman's PRLM-P approach.

These results show that the PPR-C system has lower performance in five language identification performance to Zissman's PRLM-P system. In particular, Zissman's PRLM-P five language recognition results on the 45 second utterances are much higher than those attained by the PPR-C system developed in this thesis.

English/Japanese/Spanish experiments were also performed on the PPR-C system for further comparison with Zissman's PRLM-P system and his parallel phoneme recognition (PPR) system. These results are presented in Tables 5.4 and 5.5. These results show that the PPR-C system has comparable performance with Zissman's

Table 5.2: Five Language Confusion Matrices

Ten Second Utterances Test					
Actual Language	Hypothesized 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 Second Utterances Test					
Actual Language	Hypothesized 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

Table 5.3: Five Language Identification Results

	Five-Language		Pairs Average	
	45-sec	ten-sec	45-sec	ten-sec
PRLM-P	86.5%	75.7%	94.7%	89.2%
PPR-C	71.9%	70.3%	88.0%	86.5%

Table 5.4: English/Japanese/Spanish Language Pair Identification Results

	Two Language Identification							
	English/Spanish		English/Japanese		Japanese/Spanish		Average	
	45-sec	ten-sec	45-sec	ten-sec	45-sec	ten-sec	45-sec	ten-sec
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%

Table 5.5: English/Japanese/Spanish Three Language Identification Results

	Three Language Identification	
	45 second utterances	ten second utterances
PRLM-P	92.3%	85.1%
PPR	86.5%	85.1%
PPR-C	82.7%	82.2%

PPR and PRLM-P systems on each of the three language pairs. This should be the case since the two systems are trained and tested on the same data and are using basically the same approach.

The results of three language English/Japanese/Spanish identification are given in Table 5.5. Zissman's PRLM-P system had the best results with his PPR system performing slightly below that and the PPR-C system having the worst results just slightly below Zissman's PPR system. The discrepancy between the two PPR systems could be attributed to the fact that Zissman's PPR system used the monophones plus the top 100 most commonly occurring diphones from the training data whereas the PPR-C system used only monophones, though it is not clear why this effect was not observed in the paired language case.

5.5 Conclusion

The results from the English/Japanese/Spanish experiments validate the PPR-C system since these results are comparable to those of Zissman's PPR tests. In addition

both PPR systems had comparable results with Zissman's PRLM-P system. Thus, for language identification on up to three languages, the method of using phone-based acoustic likelihoods is good and produces relatively accurate results.

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

Chapter 6

Conclusion

This thesis demonstrated that language identification on telephone speech using phone-based acoustic likelihoods is feasible, but does not yet produce results comparable to other systems. On three language identification, the PPR-C system developed in this thesis 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 Zissman's 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 multi-language speech corpora, word specific modeling approaches will be more appropriate.

Bibliography

[1] H. Dahl. *Word Frequencies of Spoken American English.* Verbatim, 1979.

[2] H. Hermansky et al. RASTA-PLP speech analysis technique. In *ICASSP '92 Proceedings*, volume 1, pages 121–124, March 1992.

[3] J.-L Gauvain and L. F. Lamel. Identification of non-linguistic speech features. In *DARPA Human Language Technology '93 Proceedings*, March 1993.

[4] J.-L Gauvain and L. F. Lamel. Identifying of non-linguistic speech features. In *Proceedings of Eurospeech 93*, volume 1, pages 23–30, 1993.

[5] F. J. Goodman, A. F. Martin, and R. E. Wohlford. Improved automatic language identification in noisy speech. In *ICASSP '89 Proceedings*, volume 1, pages 528–531, May 1989.

[6] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Proceedings of Eurospeech 93*, 1993.

[7] A. S. House and E. P. Neuburg. Toward automatic identification of the language of an utterance. I. preliminary methodological considerations. In *J. Acoust. Soc. Amer.*, September 1977.

[8] Alphonse Juilland and E. Chang-Rodriguez. *Frequency Dictionary of Spanish Words.* Mouton & Co., 1964.

[9] L. F. Lamel and J.-L. Gauvain. Cross-lingual experiments with phone recognition. In *ICASSP '93 Proceedings*, volume 2, pages 507–510, April 1993.

[10] Y. Muthusamy, K. Berkling, T. Arai, R. Cole, and E. Barnard. A comparison of approaches to automatic language identification using telephone speech. In *Proceedings of Eurospeech 93*, volume 2, pages 1307–1310, 1993.

[11] Y. K. Muthusamy and R. A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *ICSLP '92 Proceedings*, volume 2, pages 1007–1010, October 1992.

[12] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *ICSLP '92 Proceedings*, volume 2, pages 895–898, October 1992.

[13] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *ICSLP '92 Proceedings*, volume 2, pages 1011–1014, October 1992.

[14] M. Sugiyama. Automatic language recognition using acoustic features. In *ICASSP '91 Proceedings*, volume 2, pages 813–816, May 1991.

[15] M. A. Zissman. Automatic language identification using Gaussian mixture and hidden Markov models. In *ICASSP '93 Proceedings*, volume 2, pages 399–402, April 1993.

[16] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *ICASSP '94 Proceedings*, 1994.