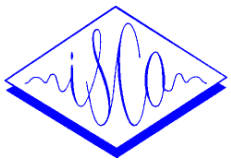

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SPONTANEOUS SPEECH LANGUAGE IDENTIFICATION WITH A KNOWLEDGE OF LINGUISTICS

Shubha Kadambe and J. L. Hieronymus
AT & T Bell Laboratories, 600 Mountain Av., Murray Hill, NJ 07974

ABSTRACT

A task independent spoken Language Identification (LID) system for telephone speech is described. This system is based on continuos density second-order ergodic variable duration Hidden Markov phoneme models and trigram phonemotactic models. The language specific phoneme models are trained using "High accuracy phoneme recognition system" [1]. A trigram phonemotactic model for each language is trained using a text corpus of about 10 million words and a grapheme to phoneme converter. The language L_i of an incoming speech signal x is hypothesized as the one that produced the highest likelihood $P(x|\beta_i)P(\beta_i|L_i)$ for all the phonemic models β_i of a given set of phonemes per language. The LID results for three languages are presented. The effect of the phonemotactic model in distinguishing languages is demonstrated by comparing the LID results obtained with and without phonemotactic models.

1. INTRODUCTION

The automatic LID system can be used to pre-sort a spoken language prior to providing telephone or computer based services such as (a) travel, (b) information (such as weather), (c) automatic translation and (d) emergency. It can also be used in national security applications. Due to the wide range of applications of LID systems, it is a research topic for the past twenty years. However, only a few studies have been reported [2] probably due to lack of common multi-language speech data base. Recently, there is revived interest in LID problem. In addition, Oregon Graduate Institute (OGI) [3] has collected twenty language speech data base.

The earlier studies [4]-[6] tried to identify languages by making use of acoustical differences in languages. However, the languages of the world differ from one another along many dimensions which have been captured in linguistic categories and therefore, the discriminative power of the LID system will not be sufficient enough to identify languages (especially when languages have similar acoustic features), if only acoustical differences are used. Hence, the current research [7, 8] is focusing on combining linguistics knowledge with speech features to identify languages.

Humans, especially linguists, have a special ability to identify a language which they have heard before, even though they are not proficient enough in it to neither understand what was being spoken nor speak that language. Some of this ability comes from generalizations about languages, consonant clusters, syllable structure, stress patterns, the prosodic features, etc. In short, humans pick out some distinguishing features of a language from a brief exposure to it and hence able to identify a language. Conceptually, our approach to LID problem is similar to how an expert in linguistics such as *Prof. Higgins of My Fair Lady* identifies a language. We would like to distinguish languages using (1) phone/phoneme inventory, (2) phonemotactics, (3) syllable structure, (4) lexical and (5) prosodic differences. In the baseline LID system that is described here, we are only making use of differences in phoneme inventory and phonemotactics to identify languages. However, in future we plan to increase the discriminative power of our baseline LID system by making use of the differences in other features mentioned above.

Our baseline LID system is similar to the LID system described in [8, 7]. In [8, 7], the language identification is achieved by using context independent phone models and unigram phonotactic constraints. However, in our system, we use context dependent phoneme models and trigram phonemotactic constraint. Since the most frequent words are monosyllabic in most languages, a trigram phonemotactic model should capture much of the phoneme sequence constraints and hence should help in discriminating languages more efficiently than unigram phonotactic constraint. The organization of this paper is as follows: in section 2, we describe our baseline LID system and the training methods that were used to train the phoneme and phonemotactic models, in section 3, we discuss the experimental results and in section 4, we conclude and indicate future plans.

2. DESCRIPTION OF THE LID SYSTEM

The block diagram of the LID system is as shown in Figure 1. In the following subsections, detailed description of each of the block in Figure 1 is given.

2.1. Phoneme recognition system

The phoneme recognition portion of the LID system is based on a high accuracy phoneme recognition system developed by A. Ljolje [1] at Bell laboratories. The structure of the phoneme recognizer is similar to a second order ergodic Continuous Variable Duration Hidden Markov Model (CVDHMM) which is also referred to as a Hidden Semi-

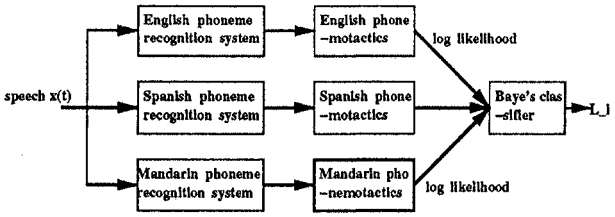

Figure 1. The block diagram of the baseline LID system

Markov Model(HSMM) in literature. The ergodic HMM has one state per phoneme. However, each phoneme is modeled by a time sequence of three probability distribution functions (pdfs) with each pdf representing the beginning, the middle and the end of a phoneme, respectively. This structure is equivalent to a three state left-to-right HM 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 95^{th} percentile of the distribution. A diagram of a Second Order Ergodic HSMM is shown in Figure 2.

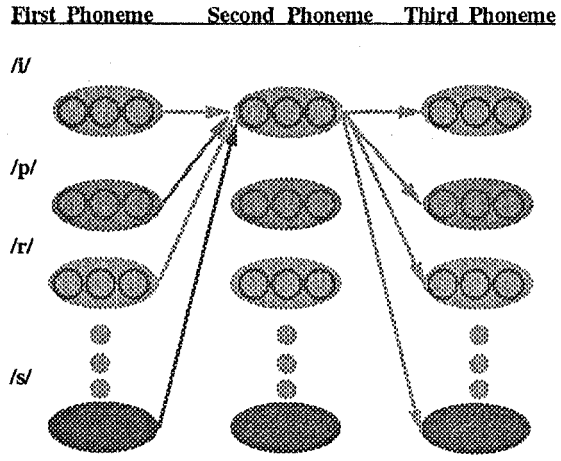

Figure 2. The architecture of a second order ergodic HSMM.

The acoustic features that are used to train the phoneme models are obtained as follows. From the telephone speech sampled at 8 kHz, 12 cepstra, 12 delta cepstra and 12 delta delta cepstra, delta energy and delta delta energy are computed using an autocorrelation LPC model with a 20 msec window and a displacement of 10 msec frame. From these 38 coefficients, 26 are chosen using a discriminant analysis method.

2.1.1. Training phoneme models:

One of the three following training procedures is used to train the phoneme recognition system depending on the type of transcription (label) and the alignment of speech waveform with the transcription is available.

The first method requires word labels and alignment of word labels with the speech waveform. From this data, the phonemic transcription and the estimated duration for each phoneme of a word label are obtained by using a text to speech (TTS) system. These durations are linearly stretched to cover the interval of the word duration. The phonemically segmented speech data thus obtained is used to initially train the HSMM models. The HSMM models are then retrained using a segmental k-means algorithm. This process is iterated until the models converge.

The second method requires time aligned phonetic transcriptions of the speech data. The initial phoneme models are trained using this labeled data. The phoneme models are retrained using a segmental k-means algorithm until the models are stable. It is possible to use these models to segment a larger corpus, and then the combined data can be used to retrain the models.

The third training method is used if only sentence level transcription and sentence level time aligned speech data is available. Similar to the first method, a TTS system is used to provide a phoneme string and durations. These are fitted to the total duration of the sentence. Initial HSMM models are trained on this data and these initial models are then used to re-segment the speech data, using the phoneme strings provided by the TTS system. Finally, the stable models are obtained by iteratively training the models using a segmental k-means algorithm similar to method 1 and 2 described above. The phonemic boundaries obtained by this procedure are less reliable than the ones obtained from the hand labels; however, the system converges to stable models. This method is similar to a flat start k-means training procedure.

2.2. Phonemotactics

In this section, we describe the block corresponding to Phonemotactics in Figure 1. For the transition probabilities of a second order ergodic HSMM, a trigram phonemotactic model is used. This provides more discriminative power than the phoneme inventory and unigram probabilities since the trigram phonemotactic capture the allowable phoneme sequence in any given language very efficiently. For example, the allowable or not allowable three phoneme sequences in English, Spanish and Mandarin are tabulated in Table 1 with trigram probability values. From this table, it is clear that the three phoneme sequence allowed in one language is not allowed in other languages.

Generally, the transition probabilities (phonemotactics) are trained using large amount of labeled speech. However, in the absence of enough transcribed speech to train the transition probabilities, they can be approximated using a large amount of text and a grapheme to phoneme converter. Therefore, in our LID system, we have trained phonemotactic models using large amount of text. Since our goal is to develop a task independent LID system, the phonemotactic models are trained using about 10 million words per language on many subjects which are obtained from the sources such as news wires, newspapers and transcribed speech.

lang \ Seq	b/oU/n(boun)	xoi (hoy)	axo (ajo)	/sr/@n(shan)	/ts/>N (zong)
English	$1.75\text{x}10^{-5}$	0.0	0.0	0.0	0.0
Spanish	0.0	$1.54\text{x}10^{-5}$	$2.33\text{x}10^{-4}$	0.0	0.0
Mandarin	0.0	0.0	0.0	$5.86\text{x}10^{-4}$	$6.05\text{x}10^{-4}$

Table 1. Allowable or not allowable trigram phoneme sequences in different languages.

2.2.1. Training phonemotactic models:

The block diagram of the system that is used to train the phonemotactics is as shown in Figure 3. In the following paragraph, we describe each block of Figure 3 in detail. A grapheme to phoneme converter is used to convert large amount of text into strings of phonemes. Since the grapheme to phoneme converter of each language uses its own phoneme symbol set, phoneme symbols outputted by each grapheme to phoneme converter is converted to worldbet phoneme symbols using the worldbet symbol converter. The worldbet phoneme symbol's representation is developed by J. Hieronymus[9] and it covers all most all the phonemes of languages of the world. Finally, the trigram probability estimator estimates the probability values using the following equation:

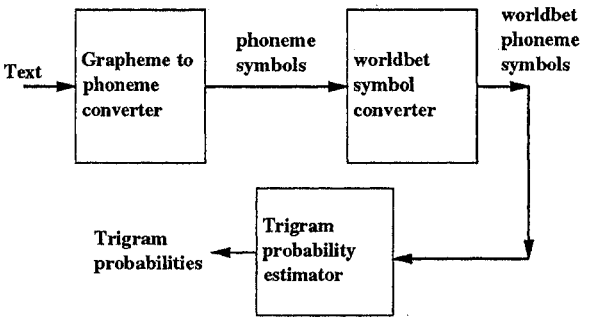

Figure 3. The block digram of the method used to train the phonemotactic model

$$\Pr(s_3|s_1,s_2) = \lambda_3 f(s_3|s_1,s_2) + \lambda_2 f(s_3|s_2) + \lambda_1 f(s_3) \quad (1)$$

where, weights λ_3, λ_2 and λ_1 are set 1, 0 and 0, respectively, s_i is the phoneme symbol i and $f()$ is the frequency of occurrence. In the next section, we describe the third block of Figure 1 namely the Baye's classifier which is used to classify an incoming speech signal into one of the languages that the LID system is trained.

2.3. Baye's classifier

For language identification, the subsystems (block 1 and 2 in Figure 1) for each language are run in parallel for a given speech signal. The language subsystem with the highest log likelihood is chosen as the language of the input speech signal. The log likelihood is computed on a per frame basis to avoid the bias toward short utterances. In addition, since the phoneme set of each language contains different number of phonemes (for example, the phoneme set of English has 42 phonemes where as Mandarin and Spanish have 41 and 27 phonemes, respectively), the computation of the log likelihood on a frame basis help in achieving the normalization with respect to the number of phonemes. The log likelihood is computed using the Bayes rule $P(x|L_i) = P(x|\beta_i)P(\beta_i|L_i)$ where the Ps are conditional probabilities, x is the input speech signal, β_i is the phoneme sequence and L_i is the phonemotactic model of the language i.

3. EXPERIMENTAL DETAILS AND RESULTS

A subset of OGI multi-language speech data base is used to train and test the three language (English, Spanish and Mandarin) identification system. In the following subsections, we discuss the experimental details of training and testing this system.

3.1. Training

3.1.1. phoneme models:

The English, Spanish and Mandarin phoneme recognition systems were trained using about 67 minutes of speech per language. This 67 minutes of speech corresponds to about 80 different speakers with speech utterance of length equal to about 50 secs per speaker. At the time of training this LID system, the word labels and the word label time alignment with speech waveform were available for English and Spanish and, time aligned phonetic transcription was available for Mandarin. Therefore, the English and Spanish phoneme recognition systems were trained using method 1 and the Mandarin system was trained using method 2. The training methods 1 and 2 are described in section 2.1.1. The phoneme models that are obtained for English and Spanish were observed to be not as good as in the case of Mandarin, since there are only about 75 % agreement between the hand labels and the labels obtained from the TTS systems. The discrepancy between the hand labels and the labels obtained from the TTS systems is because the TTS systems tend to over articulate and hence over segment. For example, the phonetic transcription of the word "and" obtained from the TTS system is always "@ n d"; however, in reality the sound "d" is very rarely produced by a speaker while speaking continuously.

3.1.2. Phonemotactic models:

The phonemotactic models for English, Spanish and Mandarin were trained using about 7.5, 6.0 and 10 million words, respectively. For English and Spanish most of these words were obtained from AP news wire after filtering the header information, stock price and other numerical data. In addition, transcribed speech from the training data was also used. In the case of Mandarin, most of the 10 million words were obtained from two news paper sources after applying the similar pre processing method described as in the case of English and Spanish. The transcribed speech from the training data was also included to train the phonemotactic model. The trigram probability values were estimated using Equation (1) after converting the text to phoneme strings as explained in section 2.2. In the case of English and Spanish, respective TTS systems [10, 11]

were used to convert text to phoneme strings where as in the case of Mandarin, Mandarin pronouncer [12] was used.

3.2. 3.2 Testing

After training the three language identification system described in the previous section, it was tested using the test data. The test data was divided into two parts. The first part consists of about 18 speakers per language. The length of speech utterance of each speaker is about 50 secs. The second set consists of about 72 chunks of 10 secs long utterances. These 10 secs long utterances were obtained by segmenting a 50 secs long utterance of each speaker of the test data into 4 segments as specified by NIST (an agency which evaluated the LID systems developed at various sites). The test results obtained using these two test data set are tabulated in Table 2. From this table we can see that we have obtained an average of 91 % LID rate on three languages when 50 secs long utterances were used and 84 % LID when 10 secs long utterances were used. Since the drop in LID rate when short segments of speech (10 secs) are used is not very significant, it implies that our LID system is efficient in identifying a language even when short segments of speech are used.

Length	English vs Spanish and Mandarin		Mandarin vs English and Spanish		Spanish vs English and Mandarin	
~50 sec	84 %		94 %		94 %	
10 sec	80 %		83 %		90 %	
	English Mandarin	Mandarin English	English Spanish	Spanish English	Mandarin Spanish	Spanish Mandarin
~50 sec	100 %	100 %	89 %	94 %	94 %	94 %
10 sec	97 %	98 %	86 %	93 %	83 %	93 %

Table 2. Test results using phonemotactics constraint

In order to test the effect of the phonemotactics on LID rate, the LID system was tested without using the phonemotactic's constraint. The test results so obtained are tabulated in Table 3. From this table, we can see that the average

Length	English vs Spanish and Mandarin		Mandarin vs English and Spanish		Spanish vs English and Mandarin	
~50 sec	74 %		66 %		76 %	
	English Mandarin	Mandarin English	English Spanish	Spanish English	Mandarin Spanish	Spanish Mandarin
~50 sec	100 %	100 %	74 %	76 %	66 %	94 %

Table 3. Test results without using phonemotactics constraint

LID rate dropped to 72 % when phonemotactics constraint was not used. This indicates that the addition of higher level linguistics knowledge such as phonemotactics help significantly in discriminating languages. In addition, we can also see that the effect of phonemotactic constraint is more in the case of languages which have similar acoustic features as evidenced by no drop in the LID rate of language pair English and Mandarin and by the significant drop of the LID rate of language pair English and Spanish.

4. CONCLUSION

In this paper, we have described a task independent spoken LID system based on linguistics knowledge. We have demonstrated that the discriminative power of the LID system can be improved by adding higher level linguistics knowledge such as phonemotactics. We have also demonstrated that the LID system described here, is efficient in identifying a language when short segments of speech is used. Future work warrants the addition of lexical access, prosody, etc. to improve the LID rate.

REFERENCES

[1] A. Ljolje, *High Accuracy Phone Recognition Using Context Clustering and Quasi-triphonic Models*, Computer Speech and Language, to appear.

[2] Y. K. Muthusamy, *A Review of Previous Work in Automatic Language Identification*, Technical report No. CS/E 92-009, Center for Spoken language understanding, Oregon Graduate Institute, February 1992.

[3] Y. K. Muthusamy, R. A. Cole and B. T. Oshika, *The OGI Multi-Language Telephone Speech Corpus*, Proc. of ICSLP 92, Banff, Canada, 1992.

[4] A. S. House and E. P. Neuberg, *Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations*, Journal of the Acoustical Society of America, Vol. 62, No. 3, pp. 708-713, 1977.

[5] D. Cimarusti and R. B. Ives, *Development of an Automatic Identification System of Spoken Languages: Phase 1*, Proc. of ICASSP 82, Paris, France, May 1982.

[6] K. P. Li and T. J. Edwards, *Statistical Models for Automatic Language Identification*, Proc. of ICASSP 80, Denver, CO, April 1980.

[7] M. A. Zissman and Elliot Singer, *Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modeling*, Proc. of ICASSP 94, Adelaide, Australia, April 1994.

[8] L. F. Lamel and J. L. Gauvain, *Language Identification Using Phone-based Acoustic Likelihoods*, Proc. of ICASSP 94, Adelaide, Australia, 1994.

[9] J. L. Hieronymus, *ASCII Phonetic Symbols for the World's Languages: Worldbet*, preprint.

[10] R. Sproat, *NewTTS: a user's manual*, Technical Memorandum, At & T Bell laboratories, TM # 11222-921102-09TM, November 1992.

[11] J. Gregorio and E. Sardina, *Descripci'on de la arquitectura del conversor texto-voz amigo*, Internal Document, Telefonica Investigaci'on y Desarrollo.

[12] R. Sproat, Chilin Shih, W. Gale and N. Chang, *A Stochastic Finite-State Word-Segmentation Algorithm for Chinese*, to appear in proc. of ACL 32nd meeting, Las Cruces, New Mexico, June 1994.