LANGUAGE IDENTIFICATION WITH PHONOLOGICAL AND LEXICAL MODELS.

Shubha Kadambe and James L Hieronymus
AT & T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT

A task independent spoken Language Identification (LID) system which uses phonological and lexical models to distinguish languages is described in this paper. We demonstrate that the performance of a LID system which is based only on acoustic models can be improved by incorporating higher level linguistic knowledge in the form of trigram phonemotactics and lexical matching. We also present the performance of our LID system for four languages (English, German, Mandarin and Spanish).

1. INTRODUCTION:

In the future, LID systems will be an integral part of telephone and speech input computer networks which provide services in many languages. A LID system can be used to pre-sort the callers into the language they speak, so that the required service will be provided in an appropriate language. Examples of these services include, travel information, emergency assistance, language interpretation, telephone information and stock quotations.

Language identification has been the subject of research for several years. Initially, systems were developed to screen radio transmissions and telephone conversations for the intelligence community. The performance of these systems was not good enough to use them for on-line service applications, which requires above 90 % correct identification of the language that is spoken.

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

Humans, especially linguists, have a special ability to 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 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 was described in [8], we only made use of differences in phoneme inventory and phonemotactics to identify languages. However, in the LID system that is described here, we improve the performance of our baseline system in particular, in the case of English and Spanish by making use of the lexical differences in these two languages. In addition, we have extended the capability of our baseline system to identify four languages.

Some early systems [3]-[5] identified languages by using broad phonetic categories and spectral characteristics around the vowel classes. In the past three years, a number of researchers [6, 7, 8] have been developing systems which first recognize phonemes and then use a phonemotactic model of phoneme sequences allowed within each language to identify the spoken language. Our baseline system described in [8] uses a continuous density, second order ergodic variable duration hidden Markov models to achieve the phoneme recognition based on tri-phonemes and trigram phonemotactic models. However, the other recent LID systems [7, 6], use context independent Markov models for phoneme recognition with bigram phonemotactic models. For languages with Consonant-Vowel-Consonant (CVC) syllable structure, the trigram models do a very good job of modeling the most frequent words, which are usually mono-syllabic and hence, should help in discriminating languages more efficiently than bigram phonemotactic models.

In the following sections, we briefly describe our baseline LID system and describe the lexical access module that is interfaced with the base line system. We present the results and finally, conclude and discuss our future goals.

2. DESCRIPTION OF THE BASE LINE LID SYSTEM:

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

2.1. Phoneme recognition system

The phoneme recognition system is based on "a high accuracy phoneme recognition system" developed by A. Ljolje [1]. This phoneme recognizer is based on a second order ergodic Continuous Variable Duration Hidden Markov Model (CVDHMM). 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

0-7803-2431-5/95 $4.00 © 1995 IEEE

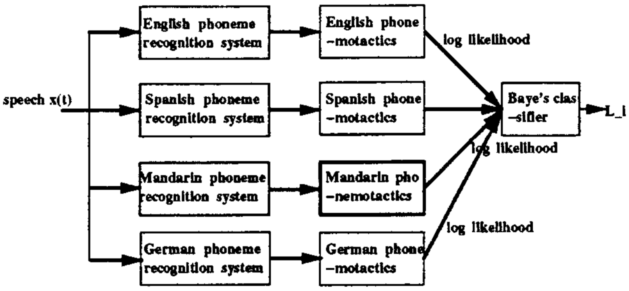
speech x(t)
English phoneme recognition system
Spanish phoneme recognition system
Mandarin phoneme recognition system
German phoneme recognition system
English phone -motactics
Spanish phone -motactics
Mandarin pho -nemotactics
German phone -motactics
log likelihood
log likelihood
log likelihood
Baye's clas -sifier
L_I

Figure 1. The block diagram of the baseline LID system

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.

2.1.1. *Training phoneme models:*

Different training procedures were adopted to train the phoneme recognition system depending on the type of transcription and the alignment of speech waveform with the transcription is available. (1) When the word labels and the alignment of these labels with the speech waveform is available, the phonemically segmented data was generated automatically by obtaining the phonemic transcription and the estimated duration for each phoneme using a Text-To-Speech (TTS) system and stretching these durations linearly to cover the word duration. The phonemically segmented data thus obtained is used to initially train the ergodic HMM models. Models are re-trained using a segmental k-means algorithm iteratively until the models converge. (2) When the time aligned phonemic transcription of the speech data is available, the initial models are trained using this data and the models are re-trained using the segmental k-means algorithm iteratively until the models converge. (3) When the sentence level transcription and segmentation is available, the phonemic level transcription and segmentation is obtained automatically as described in method 1 except that the phoneme durations are stretched linearly to fit the whole sentence. The models are trained iteratively as described in method 1 by using the segmented data so obtained. 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

For the transition probabilities of a second order ergodic HMM, a trigram phonemotactic model is used. This provides more discriminative power than the phoneme inventory and bigram 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 some of the three phoneme sequences allowed in one language is not allowed in other languages. Usually, the transition probabilities (phonemotactics) are trained using large amounts of labeled speech. However, in the absence of enough transcribed speech to train the transition probabilities, they can be approximated using a large amounts of text and a grapheme to phoneme converter. Therefore, in our LID system, we have trained phonemotactic models using large amounts 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 which are obtained from different sources such as news wire services, newspapers and transcribed speech. The trigram phonemotactic models are trained by converting text to phoneme strings and then by estimating the trigram probability values by applying the following equation.

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

Where, the weights λ_3, λ_2 and λ_1 are set to 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 Bayesian classifier which is used to classify an incoming speech signal into one of the languages that the LID system is trained.

Seq lang	b/oU/n(boun	xol (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.

2.3. Bayesian 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 Baye's 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.

We interfaced a lexical access module with the phoneme recognition system to further improve the performance of our baseline system. The block diagram of the modified LID system is as shown in figure 2. In the next section, we briefly describe the lexical access block of this figure.

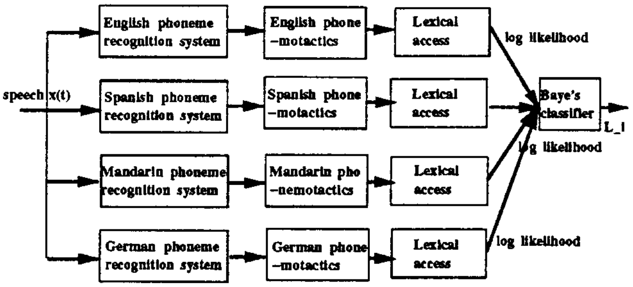

Figure 2. The block diagram of the modified LID system

3. LEXICAL ACCESS

The lexical access module in the above figure is based on the "Weighted rational transduction and their application to human language processing" [9]. This method uses the concepts of weighted language, transduction and finite state automata from algebraic automata theory to decode cascades in speech and language processing. Lexical access can be considered as a transduction cascade since the lexical access problem can be decomposed into a transduction, "D" from phoneme sequences to word sequences and a weighted language, "M", which specifies the language model. Each of these can be represented as a finite-state automaton. The automaton for the phoneme sequence to word sequence transduction "D" is defined in terms of word models. A word model is a transducer from a subsequence of phoneme labels to a specific word. To each subsequence of phonemes, a likelihood is assigned indicating that it produced the specified word. Hence, different paths through a word model correspond to different phonetic realizations of the word which has an advantage of incorporating alternate pronunciations. The language model "M" which is generally an N-gram model can be considered as a weighted finite state acceptor. Combining the automata, "D" and "M" results in an automaton which assigns a probability to each word sequence and the highest probability path that the automaton estimates gives the most likely word sequence for a given utterance. Thus, a best sequence of words which correspond to a given utterance can be obtained. For language identification, the subsystems (block 1, 2 and 3 in Figure 2) for each language are run in parallel for a given speech signal similar to the base line system described above. The language subsystem with the highest log likelihood is chosen as the language of the input speech signal.

3.1. Word and language model

The transducer "D" (lexicon or word model) and acceptor "M" (language model) were built using about 10,000 words per language which were obtained from the OGI multi-language transcribed speech data base. This contained about 2000 unique words and hence, the size of the lexicon is 2000. The bigram language model was trained using the approach described in [10].

4. EXPERIMENTAL DETAILS AND RESULTS

The German language subsystem is added to the base line system described above. The four language (English, German, Mandarin and Spanish) ID system was trained and tested using the multi-language spontaneous speech data base collected by Oregon Graduate Institute [2]. The training and test data consists of about 80 and 18 speakers, respectively, per language with length of speech utterance equal to about 50 secs per speaker. The acoustic models of English and Spanish phoneme recognition systems were trained using the method 1 and, the Mandarin and German recognition systems were trained using method 2 as described in section 2.1.1..

After training the four language identification system, it was tested using the test data. The system was also tested using short intervals of speech of 10 secs long. This test set consisted of 72 chunks of 10 secs long utterances per language. These were obtained by segmenting the 50 secs long utterance from each speaker of the test data into 4 segments as specified by NIST (which evaluates the performance of LID systems developed at various sites). We obtained an average of 88 % LID rate on four languages on the 50 secs utterances and 82 % on the 10 secs utterances. In table 4., the results for four languages including language pair identification rates are tabulated. From this table we can see that the language pair identification rate is the lowest in the case of English and Spanish.

Average four language identification is: 88 % for ~50 sec speech and 82 % for 10 sec speech.

Len	Eng vs Spa, Ger and Man	Man vs Eng, Ger and Spa	Spa vs Eng, Ger and Man	Ger vs Eng, Spa and Man
~50s	84 %	94 %	94 %	80 %
10s	78 %	81 %	86%	75 %

	Eng Man	Man Eng	Eng Spa	Spa Eng	Man Spa	Spa Man	Eng Ger	Ger Eng	Spa Ger	Ger Spa	Ger Man	Man Ger
~50s	100%	100%	86%	94%	94%	94%	100%	85%	100%	90%	95%	100%
10s	97%	98%	80%	93%	83%	93%	96%	75%	98%	77%	92%	96%

Table 2. Four language and language pair identification results.

In order to improve the performance of the identification rate for English and Spanish, a lexical access module was added to our LID system as described in section 3.. The identification results that were obtained using lexical access with bigram grammar is tabulated in table 4.. From this table, we can see that the language pair result in the case of English and Spanish improved when the lexical differences in these two language were used. This difference is not large, but improved phonemic modeling will make the word recognition improve, and retraining of the phonemic models based on phonetic hand labels is planned.

Length	English vs Spanish	Spanish vs English
50 secs	91 %	96 %
10 secs	Not Available	Not Available

Table 3. English and Spanish language identification results with lexical access.

5. CONCLUSIONS

A four language identification system based on phonological and lexical models was described. LID results for four languages were reported. From the comparison of our previous three language ID results (91 % correct) with the current four language results (88 % correct), we can see that the drop in identification rate when a new language is added is not very significant. We demonstrated that the LID rate in the case of English and Spanish can be improved by making use of the lexical differences in these two languages. This implies that the discriminatory power of the LID system can be improved by adding higher level linguistic knowledge. Future work warrants addition of lexical access module in the case of other two languages and thus improve the average 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, R. A. Cole and B. T. Oshika, *The OGI Multi-Language Telephone Speech Corpus*, Proc. of ICSLP 92, Banff, Canada, 1992.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] S. Kadambe and J. L. Hieronymus, *Spontaneous speech language identification with a knowledge of linguistics*, Proc. of ICSLP, Sept 18-22, Yokohama, Japan, 1994.

[9] F. Pereira, M. Riley and R. Sproat, *Weighted Rational transduction and their application to human language processing*, DARPA Workshop on Human Language Tech, Princeton, NJ, 1994.

[10] D. Hindle, Preprint 1994.