LANGUAGE IDENTIFICATION USING MULTIPLE KNOWLEDGE SOURCES

Eluned S. Parris and Michael J. Carey.

Ensigma Ltd, Turing House, Station Road, Chepstow, Gwent, NP6 5PB, U.K.

ABSTRACT

Language identification experiments have been carried out on language pairs taken from seven of the languages in the OGI Multi-language Telephone Speech Corpus. This builds on our previous work but introduces new techniques which are used to exploit the acoustic and phonetic differences between the languages. Subword Hidden Markov Models for the pair of languages are matched to unknown utterances resulting in three measures the acoustic match, the phoneme frequencies and frequency histograms. Each of these measures gives 80 to 90% accuracy in discriminating language pairs. However these multiple knowledge sources are also combined to give improved results. Majority decision, logistic regression and a linear classifier were compared as data fusion techniques. The linear classifier performed the best giving an average accuracy of 89 to 93% on the pairs from the seven languages.

1. INTRODUCTION

Current approaches to language identification use a number of techniques including neural nets, codebooks, Hidden Markov Models (HMMs) and prosodic features. Neural nets have been used in many of the successful language identification systems to date, for example Li [1] uses a neural net for syllabic marking and minimises speaker variation prior to language identification. Codebook techniques have had more limited success since most of the temporal information is discarded and discrimination between the languages relies only on spectral differences which are probably insufficient. HMMs have been used at both the subword and language level for language identification [2]. Usually a grammar is applied either during recognition or in post-processing to constrain output to a plausible sequence of phonemes. Some of these techniques make use of the acoustic information while others use the statistics of the language, but few approaches have combined all of the information available. This paper describes a novel approach to language identification which combines multiple knowledge sources. This builds on previous work [3] but introduces new techniques which are used to exploit the acoustic and phonetic differences between the languages. These are then combined using a number of methods to give results which are better than those given by any single technique alone.

2. DATABASE

The Oregon Graduate Institute (OGI) Multi-Language Telephone Speech Corpus [4] has been used to evaluate the techniques developed for language identification. This database was designed to support research on automatic language identification and multi-language speech recognition. Eleven languages are now available: American English, French, German, Japanese, Vietnamese, Hindi, Korean, Farsi, Mandarin Chinese, Spanish and Tamil. There are at least ninety different speakers in each of the languages with an average male to female ratio of 7:3. This ratio varies across languages, e.g. German has a ratio of 6:4 and Hindi has a ratio of 8:2. Most of the languages have some speech transcribed at a broad phonetic level, e.g. vowel, fricative, but only six of the languages have fine level transcriptions. These are available for a selection of speakers in the six languages but only the free speech files have been transcribed. The research reported in this paper was carried out on seven of the languages, American English(EN), French(FR), German(GE), Japanese(JA), Farsi(FA), Mandarin Chinese(MA) and Spanish(SP).

3. ACOUSTIC PROCESSING

3.1 Front End Processing and Model Building

The speech in the OGI database is sampled at 8kHz. This has been analysed using nineteen mel-spaced filters. The log-power outputs of the filterbank are transformed using a discrete cosine transform to give the mel cepstrum of the speech at a frame rate of 10ms. The feature vector consists of twenty-six elements; energy, twelve cepstra and the time derivatives of energy and cepstra calculated over a 50ms window.

Our approach to language identification uses a subword recogniser for each language to transcribe the speech into phonetic units. Each phoneme is represented by a three state HMM with left to right topology. Multivariate Gaussian distributions with continuous mixture densities are used to model the varying speech characteristics with separate HMMs being used for male and female speakers. The fine level transcriptions of the six annotated languages in OGI have been used to construct accurate HMMs for each phoneme. The generation of automatic fine level transcriptions was needed for the remaining five

0-7803-2431-5/95 $4.00

languages. There are two previous approaches to this problem. The first uses the phonemes from one language with annotated data, e.g. American English to segment the new language. This has the major disadvantage that the phonemes which occur in the new language and not in American English, and are therefore useful in discriminating between languages, are not being modelled separately. The second approach uses all of the phonemes from annotated languages to segment the new data. This does not generate the required set of phonemes and many redundant models are produced. To overcome these problems, a set of subword models corresponding to the correct phonemes for a new language have been assembled from models built on other languages. Phonetic knowledge is used to select the appropriate models from the closest languages, e.g. French is built from the other European languages, in particular Spanish, which is in the same language group.

3.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) first proposed by Fisher [5] is a technique used in pattern classification to provide an improved feature set. Our previous work on LDA [6] showed that significant gains in performance could be obtained when mixture densities were incorporated into IMELDA [7]. This technique has now been applied to language identification, giving similar improvements and better discrimination. LDA optimises a measure of class separability

$$J = tr(W^{-1}B)$$

where W is the pooled within class covariance matrix and B is the between class covariance matrix calculated from the class centroids. An orthogonal feature set is produced which maximises discrimination. Other approaches suffer with bias towards particular languages and normalisation is needed to compensate. LDA removes this bias and normalisation problem. LDA produces models with unit variances which reduce the distance metric used in pattern matching to a simple Euclidean. This reduces the computation and storage required to less than half. The subword recognition rate was about 40% for a typical language.

4. LANGUAGE IDENTIFICATION TECHNIQUES

Three new techniques have been developed for language identification, each one using the differences between the languages to improve performance.

4.1 Acoustic Matches

The first approach assumes that corresponding phonemes in different languages are spectrally different. The probability $p(A \mid L_i)$ of the acoustics A given a language L_i is different across languages and by using Bayes theorem the probability of a language given the acoustics $p(L_i \mid A)$ can be found.

$$p(L_i \mid A) = \frac{p(A \mid L_i)p(L_i)}{p(A)}$$

Language identification is performed by matching the models of each language in parallel to the unknown utterance and calculating $p(L_i \mid A)$ for each language. The language corresponding to the models giving the best $p(L_i \mid A)$ is identified as the language of the unknown utterance. Our approach differs from others in the calculation of $p(A \mid L_i)$. The acoustic scores for each phoneme are weighted so that the final output is biased towards the models producing the best scores for the true language. Other approaches give all phonemes equal weighting irrespective of their ability to discriminate between languages.

4.2 Phoneme Frequencies

The second approach uses the knowledge that phonemes occur with different frequencies in different languages. In our previous work on speaker identification [8], the concept of usefulness was successfully applied to weight phoneme occurrences to maximise the differences between speakers. In language identification a similar technique can be used to improve results. Using Bayesian statistics it can be shown that the contribution of each phoneme to the discrimination between classes is given by

$$p(w_k \mid L_i)\log\frac{p(w_k \mid L_i)}{p(w_k \mid L_j)}$$

where $p(w_k \mid L_i)$ is the probability of phoneme w_k occurring in language L_i, $p(w_k \mid L_j)$ is the probability of phoneme w_k occurring in the other languages. The most useful phonemes occur frequently in one language and infrequently in other languages, and also have minimal variation in occurrence between speech utterances.

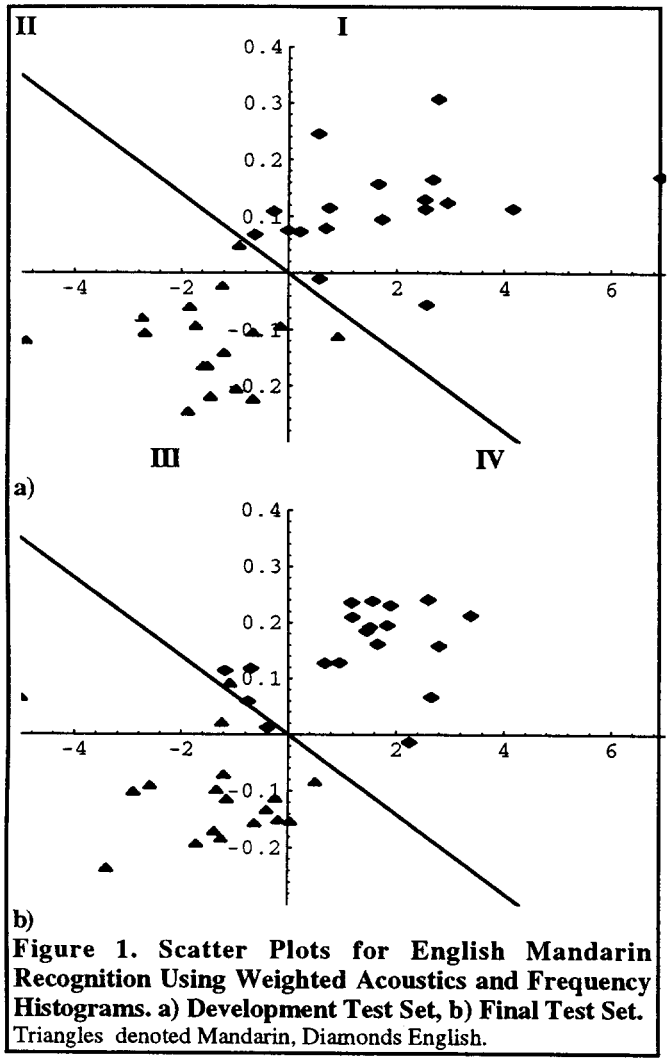

Figure 1. Scatter Plots for English Mandarin Recognition Using Weighted Acoustics and Frequency Histograms. a) Development Test Set, b) Final Test Set. Triangles denoted Mandarin, Diamonds English.

Language identification is performed by accumulating the log likelihood ratios for each set of language models. The language giving the best usefulness score is identified as the language of the unknown utterance.

4.3 Frequency Histograms.

The third approach uses the differences in distributions of all phoneme frequencies to discriminate between languages. Two reference histograms of the phoneme occurrences are produced for each language of a pair. The first histogram is produced when speech from the same language as the models is input to the recogniser. The second histogram is produced when speech from the other language, different to the models, is input to the recogniser. The histograms produced for each model set are very different and all provide useful information.

Language identification is performed by running the subword recogniser with speech from the language to be identified. A frequency histogram is produced from the phoneme output of the subword recogniser. This is then matched to each of the four reference histograms to produce correlation scores. These scores are compared and the language corresponding to the best fitting histograms is identified as the language of the unknown speech.

5. COMBINING KNOWLEDGE SOURCES

Each of the techniques described in Section 4 produce different knowledge about the languages to be identified. We observed also that classification errors given by the three techniques were uncorrelated. Hence an optimal combination of these knowledge sources should produce better results than any single technique. The simplest way to combine the sources is to use majority decision. A more sophisticated technique is logistic regression [9], a novel approach to language identification. Logistic regression is a statistical method for analysis of the relationship between an observed proportion or rate and a set of explanatory variables. It is based upon the fitting of the linear logistic model

$$\pi(\mathbf{x}) = \{1 + \exp(-\eta - \mathbf{x}^T\beta)\}^{-1}$$

where $\pi(\mathbf{x})$ is the expected value of a randomly obtained proportion $p(\mathbf{x})$ from the subpopulation corresponding to the vector $\mathbf{x} = (x_1, x_2, ...x_t)^T$ explanatory variables, η is the unknown constant term β is the vector of t regression coefficients to be estimated. In language identification there is only one entry to each subpopulation and the distribution of classes is assumed to be multivariate normal. Bayes rule can be applied to give estimates of η and β These are dependent on the means and covariance matrices of the data produced by the three techniques. Logistic regression is equivalent to a single layer perceptron (SLP), which was also tried, where the weights are the vector β. However these weights are estimated globally rather than using backpropagation.

Examination of the scatter plots shown in Figure 1 reveals that classification errors only occurred in cases where the best two techniques for a language disagreed. A new classifier, the quadrant classifier, was therefore proposed. The best two techniques are selected for a language pair and the third discarded. When the techniques agree about the language classification, that is when the classifications lie in quadrant I or III of Figure 1, the file is allocated to the correct language. The parameters of a linear classifier which optimally separates the data in quadrants II and IV are estimated and are used to classify the data in those quadrants. Since the size of the test set is small and classification errors for the best two parameters are few the two classifier parameters can be estimated manually. This technique has the advantage of focusing the parameter estimation process on the part of the test set where misclassification is likely.

Lan. %	Pho. Freq	Weig Acc.	Freq. Hist.	Maj. Dec.	Log. Reg.	Quad. Class.
EN/ GE	92/92	86/84	75/82	89/89	97/95	100/95
EN/ SP	94/92	86/89	86/81	97/97	97/94	97/100
EN/ MA	91/84	89/84	91/89	94/86	94/95	100/95
GE/ SP	86/89	77/72	91/72	89/78	91/69	100/89
GE/ MA	89/76	83/76	83/89	94/84	94/78	91/86
Ave.	90/87	85/81	85/83	93/87	95/86	98/92

Table 1 Classification Results on Selected Language Pairs.
The first figure in each entry is the % correct on the development test set, the second refers to the final test set.

	FA	FR	GE	JA	MA	SP	Ave
EN	100 85	100 92	100 95	97 92	100 95	97 100	99 93
FA		92 87	95 90	81 95	92 80	94 86	92 87
FR			89 89	89 86	94 89	100 87	94 88
GE				97 84	91 86	100 89	95 89
JA					89 83	79 83	87 87
M A						97 89	94 87
SP							95 89

Table 2 Classification Results for Pairs of Languages Using the Quadrant Classifier.
The first figure in each entry is the % correct on the development test set, the second refers to the final test set.

6. RESULTS

This section describes the results achieved on the final test section of the OGI database using the techniques described in Sections 4 and 5. These results are for the 45s story files. Table 1 shows the results for each of these techniques for a representative set of language pairs. The weighted phoneme frequencies produced the best results. Using majority decision gave only a small improvement probably because equal weight is given to all three techniques irrespective of their error performance. Neither logistic regression nor the SLP gave any improvement over majority decision. However the quadrant classifier performed well and was therefore used for a larger set of language pairs shown in Table 2. The average result for all language pairs tested was 94% on the development test set and 89% on the final test set. The results for French and Farsi were similar to those for the annotated languages indicating that the method described in Section 3 used for deriving models from other languages is effective.

7. CONCLUSIONS

This paper has introduced a series of new approaches to language identification. The use of linear discriminant analysis and phonetic knowledge for model building have improved the front end processing and reduced the computation and storage required. Acoustic weighting phonetic usefulness and histograms have all provided useful knowledge for language identification. The performance has been improved by combining the knowledge sources using a number of methods. The quadrant classifier gave the best results. The results achieved on the final test set of the OGI database are equal to or better than other recently published results using different techniques[2,10].

6. REFERENCES

[1]K P Li., 'Automatic Language Identification Using Syllabic Spectral Features.', Proc. ICASSP 1994, Adelaide, pp.297-300

[2]M A Zissman and E Singer, 'Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-gram Modeling.', Proc. ICASSP 1994, Adelaide, pp.305-308.

[3]R C F Tucker, M J Carey and E S Parris, 'Automatic Language Identification Using Sub-Word Models.', Proc. ICASSP 1994, Adelaide, pp.301-304.

[4]Y K Muthusamy, R A Cole and B T Oshika, 'The OGI Multi-Language Telephone Speech Corpus.', Proc. ICSLP 1992, Banff, pp.895-898.

[5]R A Fisher, 'The Use of Multiple Measures in Taxonomic Problems.', Contributions to Mathematical Statistics. Wiley, New York 1950, pp. 32.179 - 32.188.

[6]E S. Parris and M J Carey, 'Estimating Linear Discriminant Parameters for Continuous Density Hidden Markov Models.', Proc ICSLP 94, Yokohama, pp.215-218

[7] M. Hunt et al. 'An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination.', Proc ICASSP 91, Toronto, pp881-884.

[8]E S Parris and M J Carey, 'Discriminative Phonemes for Speaker Identification.', Proc. ICSLP 1994, Yokohama pp. 1843-1846.

[9]A Agresti,' Categorical Data Analysis.', Wiley 1990, pp. 112-119.

[10]K M Berkling, T Arai and E Barnard, 'Analysis of Phoneme-Based Features for Language Identification.', Proc. ICASSP 1994, Adelaide pp.289-292