BAYESIAN METHODS FOR LANGUAGE VERIFICATION

Eluned S. Parris[(1)], Harvey Lloyd-Thomas[(1)], Michael J. Carey[(1)] and Jerry H. Wright[(2)].

(1) Ensigma Ltd, Turing House, Station Road, Chepstow, Monmouthshire, NP6 5PB, U.K.
(2) Department of Engineering Mathematics, University of Bristol, Bristol, BS8 1TR, U.K.

ABSTRACT

This paper describes a number of techniques for language verification based on acoustic processing and n-gram language modelling. A new technique is described which uses anti-models to model the general class of languages. These models are then used to normalise the acoustic score giving a 34% reduction in the error rate of the system. An approach to automatically generate discriminative subword strings for language verification is presented. The occurrence of recurrent strings are scored using a Poisson-based significance test. It is shown that when significant sub-strings do occur in the test material they are strong indicators of the target language occurring.

1. INTRODUCTION

This paper describes work carried out on language verification, the problem of determining whether a speaker is speaking a particular language or not. Most research published to date has concentrated on language identification [1,2,3], the problem of distinguishing between two or more languages where the languages are known a priori. Language identification can be performed by running a number of verification systems in parallel and normalising across the results of each verifier to produce a language identification decision.

In a two class problem such as distinguishing between two languages, the decision process consists of simply choosing the more probable of the classes given the observed data. Difficulties arise when only one of the classes is accurately modelled and the other class is not easily modelled because the statistics are unknown or non-stationary. In language verification, the language of interest can be modelled accurately from training data but data from any language can represent the second class. A fixed threshold is generally used for decision making [4], however this does not usually work well as it is difficult to set the threshold to operate consistently at the optimum point. This paper describes a new technique using anti-models to model the second class of data, including the unknown languages.

The use of phoneme sub-strings for language verification has potential advantages over alternatives such as large vocabulary speech recognition [5]. Phoneme strings that are characteristic of a language can be learned automatically and exploited, without the need for a text transcription of training data, a comprehensive vocabulary of words or a comprehensive language model. Previous work [6] showed that language-specific features were being found using recurrent phoneme sub-strings. This paper shows how this work has been extended for language verification and presents results using the Call Friend database.

2. DATABASES

The Call Friend database has recently been collected in America for use in language recognition. The database contains twelve languages - American English, Arabic, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. There are also two dialects each of American English, Mandarin and Spanish. The speech was collected over the telephone network and taken from real conversations. There are no transcriptions available for the data.

The experiments described in this paper have been carried out using the development data taken from the National Institute of Standards and Technology (NIST) 1996 Language Recognition Evaluation. The technical objective of the evaluation is to detect the presence of a hypothesised target language given a segment of conversational speech collected over the telephone. The test segments are taken from twenty conversations for each of the target languages. There are three test durations, 3 s, 10 s and 30 s, giving a total of about 15 hours of speech.

The subword level transcriptions of the Oregon Graduate Institute (OGI) Multi-Lingual Corpus [7] have also been used in training some of the language verification systems described in this paper.

3. ACOUSTIC PROCESSING

3.1 Feature Extraction

The data was sampled at 8kHz and then filtered using a filterbank containing nineteen mel-spaced filters. The log power outputs of the filterbank were transformed into twelve static, twelve first-order and twelve second-order cepstral coefficients at a frame rate of 10ms. These coefficients were augmented by energy, first-order energy and second-order energy parameters to give a thirty nine feature vector. The mean of each of the cepstral parameters was estimated for each speech segment and subtracted from each of the feature vectors.

3.2 Model Building

Our approach to language verification uses a subword recogniser for each language to transcribe the speech into phonetic units. Each subword is represented by a three state Hidden Markov Model (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 of OGI were used to construct accurate HMMs for each subword. The Call Friend data for the same six languages was then labelled with the accurate HMMs using a subword recogniser. A second set of models was then built for the Call Friend data using these transcriptions.

Models were built for the remaining languages using a boot strapping approach previously used for language identification [2]. A set of subword models corresponding to the correct phonemes for a new language was assembled from models built on other languages. Phonetic knowledge was used to select the appropriate models from the closest languages, e.g. French was built from the other European languages, in particular Spanish, which is in the same language group. An iterative technique was then used until the model sets stabilised.

3.3 Anti-models

In language verification, the language of interest can be modelled accurately from training data but the second class of data can come from any other language, not all of which occur at training time. Using Bayes theorem, the probability or likelihood of the model given the observations is given by

$$p(m_j \mid O) = \frac{p(O \mid m_j)p(m_j)}{p(O)}$$

The prior probability of the language $p(m_j)$ is assumed to be the same for all languages and $p(O) = \sum_i p(O \mid m_i)$ giving

$$p(m_j \mid O) = \frac{p(O \mid m_j)}{\sum_i p(O \mid m_i)}$$

$\sum_i p(O \mid m_i)$ is the sum of likelihoods for all possible languages, a normalisation of $p(O|m_j)$. The exact evaluation of $p(O)$ is clearly impossible, therefore a new technique has been developed to generate a general model representing the second class of data. The general model score then normalises the score of the language being verified.

A general model was produced for each of the languages in the Call Friend database using the following approach. The subword models representing the language being verified were matched to training data taken from all the other languages in the Call Friend database. A second set of models was then built using these transcriptions. Each subword model then had an associated model called an anti-model. These anti-models were then used to model the second class of data.

3.4 Experiments

A number of experiments were carried out on the development data taken from the Call Friend database using the subword models and anti-models described above. Firstly, language verification was carried out using a subword recogniser for each target language. The acoustic score for a test file was given by the sum of best likelihoods in each frame of speech within the file. The experiments were repeated using anti-models in the recognisers for each of the target languages. In this case the acoustic score was calculated by normalising the subword score by the anti-model score.

Figure 1 shows the performance achieved by these two techniques on the 30 s test. The use of anti-models has reduced the equal error rate substantially from 47.1% to 30.9%. Similar improvements were are also made on the 3 s and 10 s tests. No normalisation has been carried out across the target languages in any of these experiments.

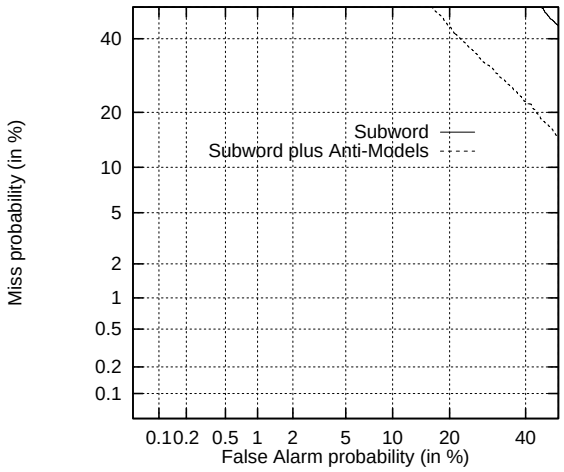

Figure 1 Comparison of Subword Models and Anti-models, 30 s Test

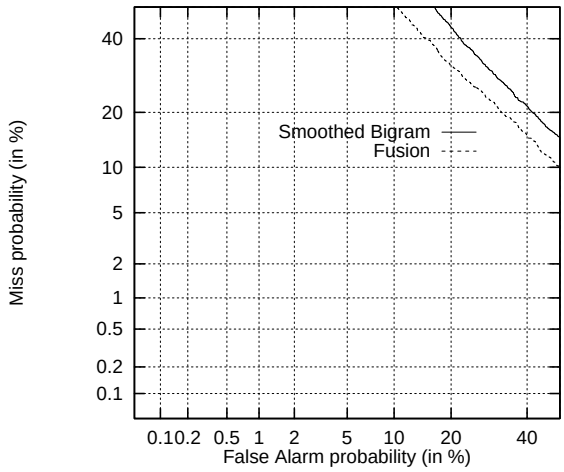

Figure 2 Comparison of Linear Interpolation Model and Fusion with Subword and Anti-models, 30 s Test

4. N-GRAM LANGUAGE MODELLING

4.1 Unigrams and Bigrams

Subword transcriptions were generated for the Call Friend training data using a subword recogniser for each of the target languages. A language model was then trained for each language from the statistics of the subwords and subword sequences output by the recogniser. The following linear interpolation model was used for the unigram and bigram statistics:

$$\tilde{P}(w_t|w_{t-1}) = \alpha P(w_t|w_{t-1}) + (1-\alpha)P(w_t)$$

where w_t and w_{t-1} are consecutive subwords observed in the recogniser output and α is a weighting factor.

At test time, the subword recogniser for a given target language was used to generate a subword transcription. The verification score was then given by the likelihood that the interpolated bigram language model produced the subword transcription.

The experiments carried out for the subword and anti-models were repeated using the linear interpolation model described above. The value of α used was 0.8, however similar performance was achieved for a range of α around this point. No normalisation was carried out across the target languages. Figure 2 shows the performance achieved for the 30 s test. The overall result is very similar to the subword and anti-model result shown in Figure 1. However, further examination of the results showed that a significant proportion of the errors between the two systems were uncorrelated. The dotted line on the graph shows the results for a simple linear weighting between the two techniques. Further improvements would be expected by using more complex data fusion techniques.

4.2 Recurrent Sub-strings

The interpolated language model described above can be extended to include higher order n-grams e.g. trigrams. However, it is difficult to estimate the parameters of the model and many of the trigrams are unseen at training time. An alternative approach is to use phoneme strings that are characteristic of the language [6]. These are learned automatically from the training data and the most discriminative strings are then used for language verification. Trigrams up to pentagrams have been used in the experiments described in this paper.

At training time, the number of occurrences of each n-gram sub-string is found for the target language. A selection of material is taken from all the other languages to represent unwanted data, and the numbers of occurrences are found for this data also. Suppose that for a sub-string there are a total of N_1 occurrences for the target language, with total file length t_1, and a total of N_2 occurrences for the remaining languages, with total file length t_2. Only the sub-strings that occur at a higher rate in the target language, i.e. $N_1 / t_1 > N_2 / t_2$ are considered. If the null hypothesis H that a sub-string occurs at the same (but unknown) rate in both data sets according to a Poisson process is adopted, then it can be shown that

$$P(N_1 \geq n_1 \cap N_2 \leq n_2 | H) \leq \sum_{k=0}^{n_2} \binom{n_1+n_2}{k} p^k (1-p)^{n_1+n_2-k}$$

where $p = t_2 / (t_1 + t_2)$.

It is easy to evaluate the right hand side of the equation from the observed counts n_1, n_2, for a particular sub-string, and the value returned can then be interpreted as the p-value for a test of significance of the hypothesis H for this sub-string. Note that no assumption of large numbers is made, and it is often the case that $n_2 = 0$. A score is associated with each sub-string equal to minus the log of this p-value, and those sub-strings are retained for which the score exceeds 2.3, corresponding to a significance level of 10%. Scores up to 15 have been seen for sub-strings that occur often in the target data and seldom or never in the remaining data. Unigrams and bigrams are excluded because their occurrences tend not to obey a Poisson process, and the test is unsuitable.

A score is generated for each test file by matching the significant sub-strings against the subword transcriptions produced by the subword recognisers for each of the target languages. At present, only exact matches are permitted. Occurrences of sub-strings can overlap, therefore the lattice of detections is parsed in order to find the highest-scoring (cumulative) path. This gives the score for the test file.

Experiments were carried out for the 30 s tests of the development data. Figure 3 shows the results achieved using recurrent sub-strings for Korean. The dotted line shows the results achieved using subword and anti-models for the same data. This figure highlights two main issues. Firstly, the use of recurrent sub-strings as a stand alone technique for language verification would not prove very successful. However, when significant sub-strings do occur they are very strong indicators for the target language and should be weighted with other techniques accordingly. Secondly, the acoustic result for Korean shown in Figure 3 is much better than the average language result shown in Figure 1. The relative performance of the techniques varies considerably across the languages. More complex data fusion techniques are currently being investigated to improve the overall performance of the system.

5. CONCLUSIONS

In this paper, two new techniques for language verification have been presented. Firstly, anti-models were used to model the general class of languages. This technique reduced the equal error rate of the system by 34% for 30 s tests, and similarly for 3 s and 10 s tests. Secondly, a new scoring technique based on recurrent sub-strings was described. The most discriminant strings for language verification were learned automatically at training time and scored using a Poisson-based significance test. Significant fragments were found to be strong indicators of a language.

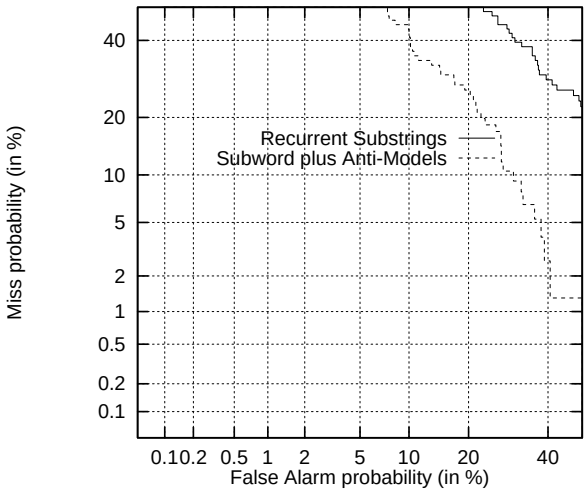

Figure 3 Comparison of Recurrent Sub-strings and Subword plus Anti-models, Korean Data, 30 s Test

6. REFERENCES

[1] M A Zissman, 'Comparison of Four Approaches to Automatic Language Identification of Telephone Speech', IEEE Trans on Speech and Audio Processing, Vol. 4, Jan 1996.

[2] E S Parris and M J Carey, 'Language Identification using Multiple Knowledge Sources', Proc ICASSP 1995, Detroit, pp 3519-3522.

[3] Y Yan and E Barnard, 'Experiments with Conversational Telephone Speech for Language Identification', Proc ICASSP 1996, Atlanta, pp 789-792.

[4] H Kwan and K Hirose, 'Unknown Language Rejection in Language Identification System.', Proc. ICSLP 1996, Philadelphia, pp 1776-1779.

[5] S Mendoza et al., 'Automatic Language Identification using Large Vocabulary Continuous Speech Recognition.', Proc ICASSP 1996, Atlanta, pp 785-788.

[6] J H Wright, M J Carey and E S Parris, 'Statistical Models for Topic Identification using Phoneme Sub-strings.', Proc. ICASSP 1996 Atlanta, pp 307-310.

[7] Y K Muthusamy et al., 'The OGI Multi-Language Telephone Speech Corpus.', Proc ICSLP 1992, Banff, pp 895-898.