LANGUAGE VERIFICATION USING ANTI-MODELS.

Eluned Parris, Harvey Lloyd-Thomas and Michael Carey.

Ensigma Ltd, Turing House, Station Road, Chepstow, NP6 5PB, U.K.
eluned@ensigma.com, harvey@ensigma.com, michael@ensigma.com

1. INTRODUCTION

This paper describes work carried out on British English language verification. Most research published to date has concentrated on language identification [1,2], the problem of distinguishing between two or more languages where the languages are known a priori. The language verification problem is to determine whether a speaker is speaking a particular language or not. This is an open set problem where any language can occur.

The language identification and verification problems are closely related and similar techniques can be used to solve both. In particular, language identification can be performed by running a number of verification systems in parallel, one for each language of interest. Normalisation across the results of each verifier can then be used 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 its statistics are unknown or non-stationary. 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. A fixed threshold does not usually work well as it is difficult to set it to operate consistently at the optimum point.

This paper describes a new technique using anti-models to model the unknown languages. Section 2 describes the databases used in this research. Sections 3 and 4 describe the speech pre-processing and model building algorithms used. Section 5 outlines the decision techniques used for language verification. Section 6 presents the results achieved using these techniques and our conclusions are given in Section 7.

2. DATABASES

2.1 Subscriber Telephony Database

The Subscriber database [3] was collected over the UK telephone network and includes over one thousand speakers from throughout the British Isles, therefore providing a good selection of data for building speaker independent models. The database consists of read speech which has been labelled at the phoneme level.

2.2 Ensigma Data Collection Exercise (ENSIGMA)

The ENSIGMA database was collected in-house over the UK telephone network. The database consists of three minute spontaneous conversations between one internal speaker and one external speaker, who have together been given a task to complete. In total, 314 conversations were collected involving 23 different male speakers. All participants were native British English speakers. In addition to the main collection, nine conversations were collected involving three different female speakers. A limited number of conversations in the database have been hand annotated at the word level and then automatically aligned at the subword level. A subset of the database was used for language verification, 104 conversation sides from all twenty three male speakers and the nine conversation sides from the three female speakers.

2.3 Translanguage English Database (TED)

The TED database is a corpus of radio microphone recordings of oral presentations given at the Eurospeech-93 conference held in Berlin. These recordings provide a large number of speakers speaking a variant of the same language (English) on specific topics. The subcorpus made available by the European Language Resources Association (ELRA) includes 188 presentations. A subset of the database was used for language verification, 75 one-minute segments from native British English speakers.

2.4 Video Mail Retrieval Database 1 (VMR1)

The VMR1 database [4] was obtained from Cambridge University and was designed and collected for use in the Video Mail Retrieval by Voice project. The database contains read training data and test data consisting of approximately twenty prompted but spontaneous spoken messages from each speaker, each of approximately one minute duration. Synchronous head and desk microphone recordings were collected in an acoustically isolated quiet room from eleven males and four females, most of whom were native British English speakers. A subset of the test data only was used for language verification, 180 messages from nine different male speakers and 77 messages from four different female speakers.

2.5 Oregon Graduate Institute Multi-Language Telephone Speech Corpus

The OGI corpus [5] was designed to support research on automatic language identification and multi-language speech recognition. The corpus contains speech in eleven languages collected over the US telephone network. These are American English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese. Each speaker spoke up to nine separate responses, ranging from single words to sixty seconds of unconstrained speech. Only the free speech files were used for language verification. The corpus has been split into three sections: train, development test and final test. For each language there are approximately fifty speakers in the training section and twenty speakers in both the development test and final test sections.

3.SPEECH PRE-PROCESSING

A speech segmentation algorithm was used to segment the data into speech and background noise. The segmentation uses a maximum likelihood pitch period estimator or periodicity detector [6] to identify regions of voiced speech. The recognition pass of the language verification is then only carried out on the parts of the data marked as speech segments.

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

4. MODEL BUILDING

4.1 British English Models

A set of forty four Hidden Markov models were built from the Subscriber database to represent the sounds of British English, the language to be verified. The models used were three state with ten mixture components per state. Each model had a left to right topology with no skipping of states allowed. The Expectation Maximisation algorithm was used for parameter estimation within the models.

4.2 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. A technique previously used in topic spotting [7] is to generate some form of general model to represent the second class of data. The approach taken for language verification was to build a subword model for each of the sounds of the language of interest, as described above. These were then matched to data from a number of languages in the OGI corpus. A second set of models was then built from these transcriptions of the OGI data. Each British English subword model then had an associated model called an anti-model. These anti-models were then used to model the second class of data.

4.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a technique used in pattern classification to provide an improved feature set [8]. LDA shows which linear combinations of features are most useful in classification and can also reduce the amount of computation and storage required by reducing the number of parameters in the system. In our previous work on language identification [2], LDA was used to improve discrimination between language pairs. For language verification the subword models and anti-models were taken together and an LDA transformation matrix produced. Each mixture component within each state of each model was treated as a separate class in the analysis. New models were produced by transforming the pooled data for each class into the new feature space and reconstructing the models directly from the data.

5. DECISION TECHNIQUES

5.1 Word-Count Scoring

In word-count scoring, two sets of models are matched to speech from an unknown language to produce the most likely sequence of models given the input. Subword models represent the language being verified and the anti-models represent any other language that may occur. It is hypothesised that the subword models will be matched in preference to the anti-models for speech from the true language, otherwise the anti-models will be matched.

The verification score S is given by

$$S = \frac{N_L}{N_L + N_{\overline{L}}} \qquad (1)$$

where N_L is the total number of matches to subword models,
$N_{\overline{L}}$ is the total number of matches to anti-models.

5.2 Frame-Likelihood Scoring

The frame-likelihood technique treats each frame of speech data independently. The probability of each state of each model is calculated for every frame of speech data. Two probabilities are stored each frame, $p(O_t \mid L)$ and $p(O_t \mid \overline{L})$, corresponding to the best subword model state and the best anti-model state. At the end of the utterance the best state probabilities are summed and the verification score S is given by

$$S = \log \frac{\sum_t p(O_t \mid L)}{\sum_t p(O_t \mid \overline{L})} \qquad (2)$$

The results using this technique can be improved by summing over the top N% of the best state probabilities. This removes poor state probabilities that occur due to noise and errors in segmentation.

5.3 Path-Likelihood Scoring

In path-likelihood scoring, language verification is performed by running two recognisers in parallel. The first recogniser uses subword models for the language of interest and the second recogniser uses anti-models to represent all other languages. The best fitting sequence of models is found for the subword models and anti-models and the probabilities compared. The verification score S is given by (2).

5.4 Usefulness

The fourth approach to language verification uses knowledge that phonemes occur with different frequencies in different languages. The subword recogniser described in 5.1 produces sequences of models from which the frequency of occurrence of each model can be calculated. Using Bayesian statistics it can be shown that the contribution of each model w_k to the discrimination between classes is given by the usefulness,

$$p(w_k \mid L) \log \frac{p(w_k \mid L)}{p(w_k \mid \overline{L})}$$

where $p(w_k \mid L)$ is the probability of model w_k occurring in language L,

$p(w_k \mid \overline{L})$ is the probability of model 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. The values of $p(w_k \mid L)$ and $p(w_k \mid \overline{L})$ are calculated from training data to include false alarms and deletions. This means that hand annotated data is not needed to calculate the probabilities. A language verification score is calculated for an utterance by accumulating the log likelihood ratios for the appropriate models.

5.5 Data Fusion

Each of the techniques described above provides different knowledge about the language to be verified. If the classification errors given by the techniques are also different then an optimal combination of any pair of these knowledge sources could produce better results than the techniques in isolation. A data fusion technique is used to combine techniques A and B and the verification score is given by

$$S = \alpha S_A + (1-\alpha)S_B \qquad (3)$$

6. RESULTS

ENSIGMA, TED, VMR1 and the final test section of OGI were used to test the decision techniques described in Section 5. Figure 1 shows the performance of the language verification system on ENSIGMA and OGI using the word-count, frame-likelihood and path-likelihood scoring techniques. Word-count scoring gives the best results with an equal error rate of 8.6%.

Figure 2 shows the difference between word-count and usefulness scoring on the TED and OGI databases. The usefulness of each model was calculated at training time from VMR1 and the development test section of OGI. Weighting the models and accumulating the log likelihood ratios gives a significant gain in performance over word-count scoring.

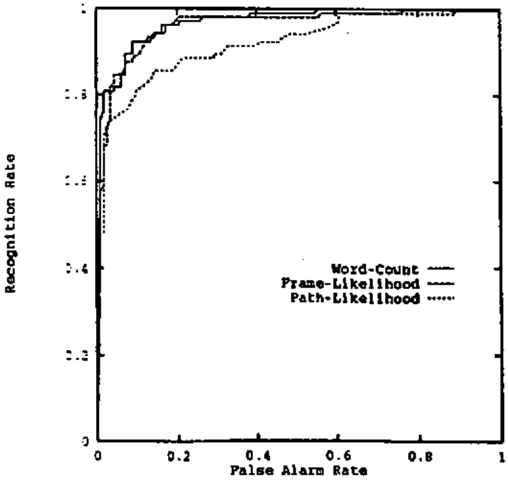

Figure 1 : Language Verification Results using ENSIGMA and OGI

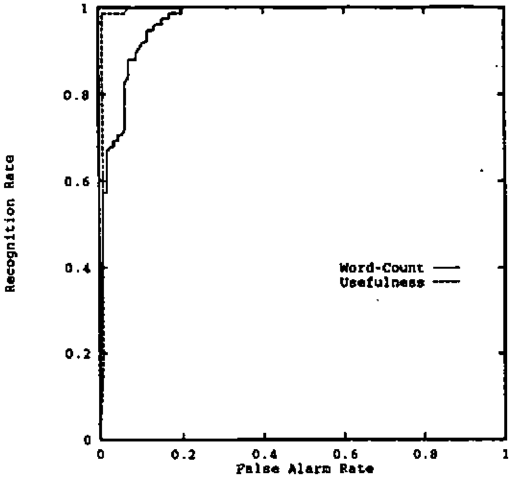

Figure 2 : Language Verification Results using TED and OGI

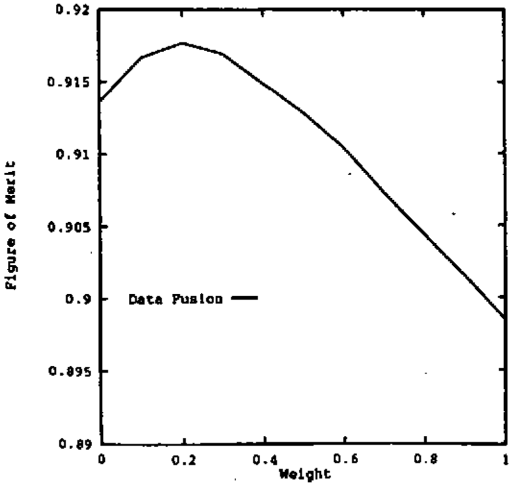

Figure 3 : Data Fusion Results using ENSIGMA and OGI

Figure 3 shows the change in the figure of merit (FOM) as the value of α in Equation (3) is varied and the word-count and path-likelihood techniques are combined. A slight improvement in FOM is achieved by using the data fusion technique on ENSIGMA and OGI.

7. CONCLUSIONS

This paper has described a number of techniques for language verification. The use of anti-models to represent the general class of data from unknown languages has been shown to work well. The word-count technique gave better results than frame-likelihood indicating that the use of sequential information is important for language verification. The best results of any single technique were achieved using the usefulness of the model classes. Combining the knowledge sources gave the best overall results with a gain in performance achieved by using one of the simplest data fusion techniques.

8.ACKNOWLEDGEMENTS

The authors would like to thank BT Laboratories for supplying the Subscriber database and Dr G. J. F. Jones of Cambridge University for supplying the VMR1 database.

9. 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.
[3] A. D. Simons and K. Edwards, *Subscriber: A Phonetically Annotated Telephony Database*, Proc. IOA (Speech and Hearing), Vol 14, part 6, Nov 1992.
[4] G. J. F. Jones et al, *Video Mail Retrieval using Voice: Report on Keyword Definition and Data Collection*, Cambridge University, April 1994.
[5] Y. K. Muthusamy et al, *The OGI Multi-Language Telephone Speech Corpus*, Proc. ICSLP 1992, Banff.
[6] D. H. Freidmann, *Pseudo-Maximum-Likelihood Speech Pitch Extraction*, IEEE Trans on Acoustics, Speech and Signal Processing, 1978.
[7] M. J. Carey and E. S. Parris, *Topic Spotting with Task Independent Models*, Proc. Eurospeech 1995, Madrid.
[8] R. A. Fisher, *The Use of Multiple Measures in Taxonomic Problems*, Contributions to Mathematical Statistics, Wiley, New York 1950, pp. 32.179 - 32.188.