RECURRENT SUBSTRINGS AND DATA FUSION FOR LANGUAGE RECOGNITION

Harvey Lloyd-Thomas†, Eluned S. Parris† and Jerry H. Wright‡

†Ensigma Ltd., Turing House, Station Road, Chepstow, NP6 5PB, UK.
‡AT&T Labs Research, Florham Park, NJ.

harvey@ensigma.com eluned@ensigma.com jwright@research.att.com

ABSTRACT

Recurrent phone substrings that are characteristic of a language are a promising technique for language recognition. In previous work on language recognition, building anti-models to normalise the scores from acoustic phone models for target languages, has been shown to reduce the Equal Error Rate (ERR) by a third. Recurrent substrings and anti-models have now been applied alongside three other techniques (bigrams, usefulness and frequency histograms) to the NIST 1996 Language Recognition Evaluation, using data from the CALLFRIEND and OGI databases for training. By fusing scores from the different techniques using a multi-layer perceptron the ERR on the NIST data can be reduced further.

1. INTRODUCTION

Our language recognition system consists of five techniques, including recurrent substrings [1], whose outputs are then fused. The following sections describe the theory behind each of these techniques and then describe the experiments and results obtained.

2. THEORY

2.1 Acoustic Models

Our language recognition system uses separate phone recognisers for each target language. Each acoustic phone model, within each recogniser, is a three state Hidden Markov Model (HMM) with left to right topology. Multivariate Gaussian distributions with continuous mixture densities are used and separate models are built for male and female speakers. Where language recognition is implemented by verification of multiple hypotheses, there are two classes of data. The first class is target languages which can be modelled accurately during training. However, the second class can come from any other language and can not be modelled explicitly. A general model is therefore built, using Bayes theorem, to normalise the score of a language being verified [2]. A general model is built for each target language using the following approach. The acoustic models representing a target language are used to transcribe training data taken from other languages. A second set of models is then built from these transcriptions. Each acoustic model then has associated with it a new model called its anti-model. These anti-models are used to model the second class of data. Linear Discriminant Analysis (LDA) is then used to generate final sets of acoustic models and anti-models for each target language.

2.2 Bigrams

Bigram models are built from the transcriptions produced by the phone recognisers for each target language. For each recogniser separate bigram models are built from training data taken from the current language of interest and also from each of the other languages of interest. The bigram statistics are smoothed by using a linear interpolation of unigram and bigram counts, similar to that used in [3].

2.3 Recurrent Substrings

Recurrent phone substrings that are characteristic of a language are generated recursively by growing shorter substrings and testing for significance. The test compares the rates of occurrence of a substring in training data from a target language and a sample of other languages. A substring may be a rare event but of high utility whenever it occurs, so a Poisson significance test which is valid even for small counts is used. A score inferred from the test statistic is associated with each substring and used for language verification, differing from the approaches taken in [4] and [5]. We consider longer n-grams (trigrams up to pentagrams) that are assumed to occur relatively infrequently and independently of each other, so that they can be modelled as a Poisson process with an occurrence rate λ. Unigrams and bigrams are excluded because their occurrences tend not to obey a Poisson process. The significance test used is based on the hypothesis that the occurrence rate λ is the same for two streams of data (a target language and a sample of other languages) but is unknown. Substrings that are highly significant by this test are then used for language verification. The test is similar to one described in [6]. Suppose for a particular substring there are N_1 (random variable) occurrences in the training data for a target language, with total length t_1 and N_2 occurrences for other languages, with total length t_2.

t_1 and t_2 are actually gross phone counts but could alternatively be durations in seconds. Assume a higher observed rate in the target language

$$n_1/t_1 > n_2/t_2$$

Also we adopt the hypothesis H that occurrences are Poisson with a common rate λ

$$N_1 \sim P(\lambda t_1) \qquad N_2 \sim P(\lambda t_2)$$

Using the following general relationship [7] between a Poisson random variable

$$N_\mu \sim P(\mu)$$

and a chi-square random variable

$$Y_\nu \sim \chi^2_\nu$$

that

$$P(N_\lambda \le x) = P\left(Y_{2(1+x)} > 2\lambda\right)$$

we have

$$P(N_1 \ge n_1 | H) = P\left(Y_{2n_1} < 2\lambda t_1\right)$$

$$P(N_2 \le n_2 | H) = P\left(Y_{2(n_2+1)} > 2\lambda t_2\right)$$

Because N_1 and N_2 are independent we infer

$$P(N_1 \ge n_1 \cap N_2 \le n_2 | H) = P\left(Y_{2n_1} < 2\lambda t_1 \cap Y_{2(n_2+1)} > 2\lambda t_2\right)$$

It follows that

$$P(N_1 \ge n_1 \cap N_2 \le n_2 | H) \le P\left(\frac{Y_{2(n_2+1)}}{Y_{2n_1}} > \frac{t_2}{t_1}\right)$$

$$= P\left(\frac{Y_{2(n_2+1)}/2(n_2+1)}{Y_{2n_1}/2n_1} > \frac{t_2/2(n_2+1)}{t_1/2n_1}\right)$$

$$= P\left(F_{2(n_2+1),2n_1} > \frac{n_1 t_2}{(n_2+1)t_1}\right)$$

where we define the F-distributed random variable

$$F_{2(n_2+1),2n_1} = \frac{Y_{2(n_2+1)}/2(n_2+1)}{Y_{2n_1}/2n_1}$$

In general, for an F-distributed random variable $F_{\nu,\omega}$ and a binominal random variable $B_{n,p}$ we have the following relationship [7] (provided ν, ω are both even numbers)

$$P\left(F_{\nu,\omega} > \frac{\omega p}{\nu(1-p)}\right) = P\left(B_{\frac{1}{2}(\nu+\omega-2),p} \le \frac{1}{2}(\nu-2)\right)$$

Setting

$$\nu = 2(n_2+1) \qquad \omega = 2n_1$$

it follows that

$$P(N_1 \ge n_1 \cap N_2 \le n_2 | H) \le P\left(B_{n_1+n_2,p} \le n_2\right)$$

$$= \sum_{k=0}^{n_2} \binom{n_1+n_2}{k} p^k (1-p)^{n_1+n_2-k}$$

where

$$\frac{\omega p}{\nu(1-p)} = \frac{2n_1 p}{2(n_2+1)(1-p)} = \frac{n_1 t_2}{(n_2+1)t_1}$$

from which

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

It is easy to evaluate the right hand side from the observed counts n_1, n_2 and total lengths t_1, t_2 for a particular substring and the value returned can then be interpreted as the P-value (actually an upper bound on it) for a test of significance of the hypothesis H for this substring. A small value indicates that the rate of occurrence for the target language is significantly higher than for the other languages. No assumption of large numbers is made and it is often the case that $n_2 = 0$. A score is associated with each substring equal to minus the log of this P-value, and those substrings are retained for which the score exceeds 2.3, corresponding to a significance level of 10%. Scores up to 15 have been seen for substrings that occur often in target data and seldom or never in the remaining data. A set of significant substrings is generated automatically for each target language by recursively growing shorter significant substrings into longer ones and counting and testing those in turn. A score is generated for each test utterance by matching the significant substrings against the phone transcriptions produced by the recognisers for each target language. At present, only exact matches are permitted. Occurrences of substrings can overlap, therefore the lattice of detections is parsed in order to find the highest-scoring (cumulative) path. This result is divided by the total number of phones to give an overall score.

2.4 Other Techniques

Usefulness This technique uses the knowledge that phones occur with different frequencies in different languages. The phone recognisers for each target language produce transcriptions from which the frequency of occurrence of each phone can be calculated. These frequencies differ when the

same phone occurs in different languages. So the frequency of occurrence of a phone differs when the true language is used as input to a recogniser, to the frequency when another language is used. These differences can be used to identify which language is being spoken.

Frequency histograms This technique uses the relative frequency of occurrence of the phones of a language to identify that language. The mean rate of occurrence of each phone can be calculated from the transcriptions produced by the phone recognisers for each language. The variability in the frequency of occurrence of each phone can also be used, provided that enough training data is available. A correlation measure is then used to compare frequency histograms of phone occurrences between training and test data.

3. EXPERIMENTS

3.1 Databases

The experiments described have been carried out using data from the NIST 1996 Language Recognition Evaluation, in accordance with the rules set out in the evaluation plan [8]. The technical objective of the evaluation was to detect the presence of a hypothesised target language, given a segment of conversational speech collected over the telephone, where the target languages were: American English, Arabic, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The CALLFRIEND and OGI databases were used to train our language recognition system as follows. Accurate HMMs for each phone were built for the six languages of OGI for which there exist phone level annotations. The CALLFRIEND data (for which no annotations existed) for the same six languages was then transcribed using these accurate HMMs. A second set of models was then built using both the CALLFRIEND and OGI data. Models were built for the remaining six languages using a bootstrap approach. A set of models corresponding to the correct phones for a new language was assembled from the models built for the original six languages. Phonetic knowledge was used to select appropriate models. An iterative process was then used to transcribe the unannotated data and build models, until the model sets produced stabilised.

3.2 Feature Extraction

Speech is sampled at 8kHz and then filtered by a filterbank of nineteen mel-spaced filters. The log power outputs from the filterbank are transformed into twelve cepstral coefficients, plus twelve first-order and twelve second-order cepstral coefficients. The cepstral coefficients are augmented by energy, plus first-order and second-order energy parameters to give a frame of thirty-nine features every 10ms. A speech segmentation algorithm is then used to identify speech segments and to discard regions of silence or noise. Finally, cepstral mean subtraction is applied to each speech segment.

3.3 System Overview

Our system consists of twelve phone recognisers, one for each of the target languages in the NIST data. Each generates acoustic likelihood scores and phone transcriptions. The system is trained by transcribing data from each target language with each recogniser, the transcriptions produced are then used to build bigram, substring, usefulness and histogram models, for each recogniser/language pair. For a given test utterance, each of the target languages is hypothesised in turn, generating a score for each of the four techniques above plus acoustics, for each recogniser/hypothesis pair. So for each test utterance a total of sixty scores are generated These scores can then be fused at different levels.

3.4 Data Fusion

Normalisation within techniques A separate multi-layer perceptron (MLP) with one hidden layer was trained independently for each technique. Each MLP had twelve inputs, corresponding to the scores for the twelve hypothesised languages for a particular technique. Twelve hidden nodes and twelve output nodes corresponding to the twelve languages to be verified were used. Independent DET curves [9] for each technique after normalisation are plotted in Figure 1. Normalising across hypotheses typically reduced the EER for an independent technique by a half. Figure 1 shows some techniques are clearly better than others, but our hypothesis is that the different techniques give rise to errors which are uncorrelated. Fusing the results from the separate techniques should therefore give better overall performance.

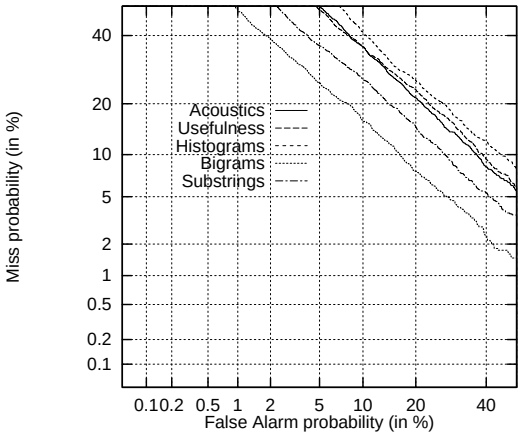

Figure 1 DET curves showing independent techniques normalised across all twelve hypotheses.

Fusion across techniques A single MLP with one hidden layer was trained for all techniques. The MLP had sixty inputs, corresponding to the scores for the twelve hypothesised languages for each of the five techniques. Sixty hidden nodes and twelve output nodes were used. A DET curve for the final system is plotted in Figure 2 and results are broken down per language in Table 1. When fusing multiple techniques we can reduce the overall EER compared to the best individual technique with normalisation across hypotheses.

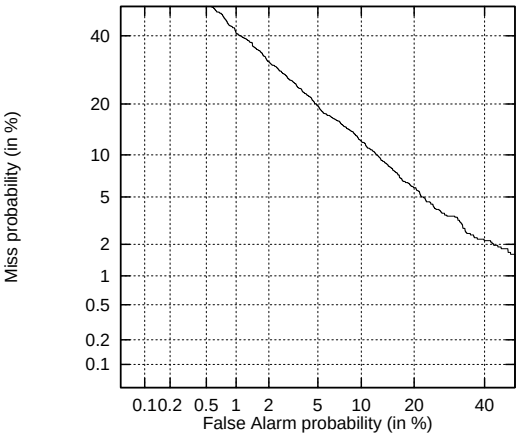

Figure 2 DET curve showing data fusion across all five techniques and all twelve hypothesis.

	% FOM	% EER
American	99.2	4.7
Arabic	95.6	9.6
Farsi	94.2	12.5
French	95.0	10.0
German	96.7	8.8
Hindi	89.3	17.2
Japanese	96.0	10.4
Korean	90.9	17.9
Mandarin	96.5	8.6
Spanish	95.1	12.4
Tamil	93.1	12.5
Vietnamese	94.2	11.4

Table 1 Per language results for data fusion across all five techniques and all twelve hypothesis.

4. CONCLUSIONS

Acoustic modelling using anti-models and LDA gives useful improvements in performance. While normalisation across hypotheses is essential to gain the most from a single technique. Data fusion across techniques gives the best system and further investigation showed that the contribution of some techniques was marginal and better results were achieved by fusing together just four techniques and not using the frequency histogram scores. We also found that fusing just bigram and substring scores gave better results than either in isolation. Although bigrams may appear to be a subset of the substring approach, the techniques model different information about the target languages. Finally, we note that results vary quite widely across languages, possibly due to poor acoustic modelling in some cases.

5. REFERENCES

1. Wright, J.H., Carey, M.J. and Parris, E.S., "Statistical Models for Topic Identification using Phoneme Substrings", *Proc. ICASSP*, pp. 307-310, May 1996.

2. Parris, E.S., Lloyd-Thomas, H., Carey, M.J. and Wright, J.H., "Bayesian Methods for Language Verification", *Proc. EUROSPEECH*, pp. 59-62, September 1997.

3. Zissman, M.A., "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44, January 1996.

4. Nowell, P. and Moore, R.K., "The Application of Dynamic Programming Techniques to Non-Word based Topic Spotting", *Proc. EUROSPEECH*, pp. 1355-1358, September 1995.

5. Lund, M.A. and Gish, H., "Two Novel Language Model Estimation Techniques for Statistical Language Identification", *Proc. EUROSPEECH*, pp. 1363-1366, September 1995.

6. Kanji, G.K., *100 Statistical Tests*, SAGE Publications, 1993.

7. Evans, M., Hastings, N. and Peacock, B., *Statistical Distributions*, John Wiley and Sons, 2nd ed., 1993.

8. NIST, "The 1996 Language Recognition Evaluation Plan", May 1996.

9. Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M., "The DET Curve in Assessment of Detection Task Performance", *Proc. EUROSPEECH*, pp. 1895-1898, September 1997.