MULTILATERAL TECHNIQUES FOR SPEAKER RECOGNITION

Eluned S. Parris and Michael J. Carey

Ensigma Ltd., Turing House, Station Road, Chepstow, Monmouthshire, U.K.
eluned@ensigma.com, michael@ensigma.com

ABSTRACT

Speaker recognition is usually accomplished by building a set of models from speech of a known speaker, training data, and subsequently using a pattern matching algorithm to score the speech from an unknown speaker, test data. In this paper we discard the notion of train and test data in speaker recognition and introduce the multilateral scoring technique. This technique comprises building speaker models on material for the known speaker and matching the unknown speaker data to these models, the traditional approach to speaker recognition. The resultant scores are fused with an equivalent set of scores produced by matching the known speaker utterance to models built on the unknown speaker data. Significant improvements have been achieved using this technique on the NIST 1996, 1997 and 1998 Speaker Recognition Evaluation data. Results are presented for two speaker recognition systems, the first based on Hidden Markov models and the second based on Gaussian Mixture models.

1. INTRODUCTION

Speaker recognition is usually accomplished by building a set of models from speech of a known speaker, training data, and subsequently using a pattern matching algorithm to score the speech from an unknown speaker, test data. In this paper we discard the notion of train and test data in speaker recognition and state the problem as follows:

"Given a sample of speech from a known speaker and a second sample from an unknown speaker determine the probability that both samples were spoken by the same speaker."

We assume that we can periodically extract a sequence of observations or feature vectors (O) from the speech which contain information about the speakers identity either in the spectral envelope, usually represented as cepstral coefficients, and/or the spectral fine structure, that is pitch. We must then estimate the likelihood that the training observations from the known speaker and the test observations from the unknown speaker are samples of the same distribution. Three approaches to this problem suggest themselves; these are,

Model Likelihoods, this is the usual approach where a parametric model is estimated from the training data and a pattern matching algorithm is used to match the test observations to the model, examples of this are Hidden Markov Models (HMMs) [1][2], Gaussian Mixture Models (GMMs) [3] or Neural Networks (NNs) [4].

Nearest Neighbours, an alternative to model likelihoods [5], is to use non-parametric techniques by computing distances using the training data directly.

Parametric Distance, comparisons are made between the distributions in the model space, models are built from both the training and test data and a distance between the models is computed [6].

In the following first we review the model likelihood approach. Then we develop the idea of multilateral scoring firstly by combining recognition scores from test speech matched to models built on the training data and then from training speech matched to models built on test data.

2. MODEL LIKELIHOODS

Unilateral Scoring. This technique draws heavily on the approach used in speech recognition in which we attempt to explain a sequence of observations by a sequence of models. In speech recognition the sequence of models is the required output of the system, in speaker recognition it is the probability of the models given the observations, $p(m_j|O)$ which is equated to the probability of the speaker $p(s)$. However typical pattern matching techniques estimate the probability or likelihood of the model given the observations. These are related by Bayes theorem,

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

Usually the prior probability of the speaker $p(m_j)$ is assumed to be the same for all speakers and is disregarded. However the estimation of $p(O)$ has led to problems in the past. Noting that $p(O) = \sum_i p(O|m_i)$ we have,

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

Now $\sum_i p(O|m_i)$ is the sum of the likelihoods for all possible speakers which normalises the likelihood $p(O|m_j)$. Hence the exact evaluation of $p(O)$ is clearly impossible. Therefore two approximations to this have been proposed. The first relies on

the observation that $p(O|m_j)$ will be small for all but a small set of similar speakers and that the approximation

$$\sum_i p(O|m_i) \approx \sum_{i \in A} p(O|m_i)$$

can be made. The set A is referred to as the cohort of speaker j and the verification score is modified by the cohort score[7]. The other approach is to construct a world or general model M which may for example be a speaker independent model in the case of a Hidden Markov Model system [8]. Then we have

$$p(m_j|O) = \frac{p(O|m_j)}{p(O|M)}$$

that is the world model score normalises the speaker's score. This has been found to work well in practice and to perform better than the same systems using cohorts.

The implicit assumption in the foregoing is that the models are derived from the training data and that the observations are feature vectors extracted from the test speech. This naturally follows from the speech recognition paradigm and is used even when no model sequence exists for example when using a GMM or Neural Network. In fact building models from test data and matching the training data to these is equally valid. The only limitation is the practical issue of having a sufficient amount of data to train the models in each case. We will refer to either of these methods as examples of *unilateral scoring*. Now we consider how we can use the information available in the training and test data more effectively.

3. CROSS VALIDATION

3.1. Bilateral Scoring

Informal observation of the performance of speaker verification systems over the years has led us to conclude that when using unilateral scoring, impostors are not reciprocal. That is when the models built from speech from speaker A are matched to speech from speaker B they do not give high likelihoods although speech from speaker A matched to the models built from speech from speaker B give high likelihoods. Assuming that samples of speech from the same speaker always matches well irrespective of which is used to build the models we have a mechanism for improving the rejection of false alarms. To do this we redefine $p(s)$ as,

$$p(s) = p(m_k|O_u, m_u|O_k)$$

where the subscripts k and u indicate the known and unknown speakers respectively. Assuming that $p(m_k|O_u)$ and $p(m_u|O_k)$ are statistically independent we have

$$p(s) = \frac{p(O_u|m_k)}{p(O_u|M)} \frac{p(O_k|m_u)}{p(O_k|M)}$$

which can be achieved by adding the log likelihood scores for the two tests. These scores may have been derived from a HMM, GMM, NN or some other pattern matcher.

3.2. Multilateral Scoring

In many cases more than one sample of data is available for the known and unknown speaker. For example, the NIST Evaluations contain a two-session training condition where the data for the known speaker has been collected on two separate occasions. Traditionally one set of models is built combining the data from all of the sessions. However, the bilateral scoring technique described above can be extended to the multilateral case to exploit the additional information available from the different sessions. Separate models are built for each session of data from the known and unknown speakers and corresponding scores produced using the pattern matching technique. The multiple scores can then be combined using a data fusion technique.

4. EXPERIMENTAL CONFIGURATION

The experiments described in the following section were carried out on the NIST 1998 Evaluation data. While several different tests were specified in that evaluation the results given here are for the 30s test conditions. The 30s test conditions give enough speech to build the second set of models on the test data. There are three 30s test conditions, one-session train, two-session train and two-session full train. There are 250 men and 250 women target speakers and a total of 2500 tests for each of the conditions.

The acoustic analysis used in the experiments was as follows. The data was sampled at 8kHz and was then filtered using a filterbank containing nineteen filters. The log power outputs of the filterbank were transformed into twelve cepstral coefficients and their first and second derivatives at a frame rate of 10ms. These coefficients were augmented by energy and delta energy parameters to give a thirty-nine element feature vector. The mean of each of the cepstral parameters was estimated for each segment of speech and subtracted from each of the feature vectors.

4.1 Hidden Markov Models

The subword models used were three state Hidden Markov Models with continuous mixture distributions and a left to right topology. No skipping of states was allowed. Speaker independent models were built as follows. A set of subword models corresponding to the forty-one phonemes of American English was built using the American-English part of the OGI Multilingual Corpus. Recognition was then performed on the training material and the results of the recognition were

compared with the annotation files to give a confusion matrix between the subword models.

The number of subword classes was then reduced by combining subword units likely to be confused, reducing the number of classes to twenty-eight. The trained classes so combined were then used to build a new set of speaker independent models including additional Switchboard data taken from the 1995, 1996 and 1997 Evaluations. Each model state had three Gaussian mixture modes.

At training time these speaker independent models were used to segment the training speech for each of the target speakers and speaker dependent models were then built from this speech using the mean estimation model building technique [1]. Each of the speaker dependent models had three modes per state. A similar set of models was also built from each of the test files.

4.2 Gaussian Mixture Models

A second speaker recognition system was set up using Gaussian mixture models (GMMs) [3]. A single speaker independent GMM was trained using the EM algorithm to maximise the likelihood of the data given the models. The GMM had 256 modes and was trained using the same data as the speaker independent HMMs described above.

Target speaker dependent GMMs were built by adapting the speaker independent GMM. Unsupervised Bayesian adaptation was used to train the speaker dependent GMMs using the training data of the target speaker. The variances of the speaker dependent GMMs were not adapted remaining identical to the speaker independent variances. This is the same technique described above for building the HMM models and proved more robust and successful than any technique where the speaker dependent variances were adapted or estimated directly from the data.

5. EXPERIMENTS

5.1 Bilateral Scoring

During recognition an unknown speaker's speech was matched to a set of Hidden Markov models comprising each of the hypothesised target speaker's dependent models and a set of speaker independent models. A score was generated for each of the target speakers, which was the percentage of the total matches achieved by that speaker's models. The z-norm technique [3] was used to align the scores across speakers. The scores were then used to generate a Receiver Operating Characteristic (ROC) which is shown in the Detection Error Trade-off (DET) [9] curve of Figure 1 as the train-test condition. The results shown are for the 1998 NIST Evaluation data, one session training condition and 30 second test. The second plot on Figure 1 shows the result of matching the training speech to the models built on the test speech, test-train.

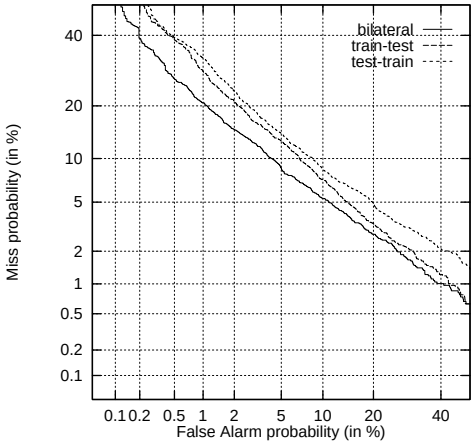

Figure 1: HMM Bilateral Results for NIST1998 Evaluation Data, One Session Training, 30 Second Test.

The third plot of Figure 1 shows the effect of fusing the train-test and test-train scores, bilateral scoring. A linear discriminant formed by adding the scores was found to be a near optimal discriminator for the 1996 Evaluation data and was therefore used for the 1998 tests. A significant improvement in performance has been achieved for all parts of the DET curve. At a 20% miss rate the false alarm probability has been reduced by a half. The equal error rate has also been reduced from 8.5% to 6.9%.

The one session training, 30 second tests described above were repeated using the 256 mode Gaussian mixture model system. During recognition an unknown speaker's speech was matched to a GMM built from the training data of the hypothesised target speaker and a speaker independent GMM. A score was generated for each of the target speakers by accumulating the log likelihood ratios for each frame. The z-norm technique [3] was used to align the scores across speakers. Figure 2 shows the DET curve achieved for the train-test condition. The second plot on Figure 2 shows the result of matching the training speech to the GMM built on the test speech, test-train.

The third plot of Figure 2 shows the effect of fusing the train-test and test-train scores, bilateral scoring. A linear discriminant was used by simply adding the scores. On the important part of the curve, a useful improvement in performance has been achieved. At a 20% miss rate the false alarm probability has been reduced by 40%. The equal error rate has also been reduced from 8.0% to 6.5%. It is expected that these results could be improved further by using GMMs with more modes e.g. 1024 or 2048. However, these results demonstrate that bilateral scoring is equally applicable to Hidden Markov models and Gaussian mixture models.

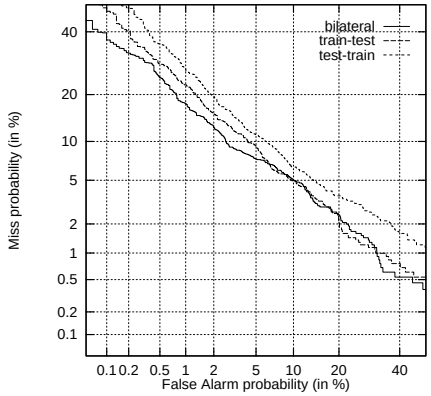

Figure 2: GMM Bilateral Results for NIST1998 Evaluation Data, One Session Training, 30 Second Test.

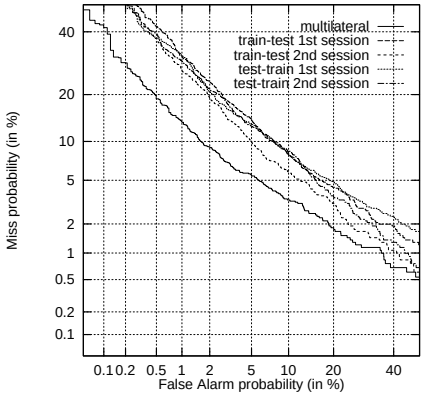

Figure 3: HMM Multilateral Results for NIST1998 Evaluation Data, Two Session Training, 30 Second Test.

5.1 Multilateral Scoring

Tests on multilateral scoring have been carried out for the two-session test condition of the NIST 1998 Evaluation using the HMM system described in 4.1. During training, separate target models were built for each of the two one-minute training sessions provided. During recognition two scores were generated for each of the two target speaker models and the z-norm technique applied. The results are presented in Figure 3 as the train-test 1st session and 2nd session DET curves. The third and fourth plots on Figure 3 show the results of matching the two separate training speech files to the models built on the test speech, test-train 1st session and 2nd session.

The third plot shows the effect of multilateral scoring, the fusing of the four train-test and test-train scores. A linear discriminant was formed by adding the four scores. A significant improvement in performance has been achieved for all parts of the DET curve. The multilateral equal error rate is 5.3% compared to 6.3% for the unlilateral train-test result using both sessions to train. The results achieved using the multilateral scoring have also been compared to the bilateral scoring technique on the same data. Currently, multilateral scoring performs only marginally better than bilateral scoring. It is expected that further improvements can be made to the multilateral scoring technique by using more sophisticated data fusion algorithms.

6. CONCLUSIONS

In this paper we have discarded the notion of train and test data in speaker recognition and introduced the idea of multilateral scoring. This technique has given significant improvements in performance on the NIST 1998 Speaker Recognition Evaluation 30 second tests. We have shown that the multilateral scoring technique is applicable to speaker recognition systems based on Hidden Markov models and Gaussian Mixture models.

7. ACKNOWLEDGEMENTS

The authors would like to thank Harvey Lloyd-Thomas and Roland Auckenthaler for their help with the speaker recognition experiments.

8. REFERENCES

1. M. J. Carey et al. "*A Comparison of Model Estimation Techniques For Speaker Verification*". Proc. ICASSP 1997 Munich, pp 1083-1086.

2. J. Baker et al. "*Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification*". Proc. ICASSP 1993 Minneapolis, pp 471-474.

3. D. A. Reynolds. "*Speaker Identification and Verification Using Gaussian Mixture Speaker Models*". Speech Comm., Aug 1995, pp 91-108.

4. J. Oglesby and J.Mason. "*Radial Basis Function Networks for Speaker Recognition*". Proc. ICASSP 1991 Toronto, pp 393-396.

5. A. Higgins et al. "*Voice Identification using Nearest Neighbor Distance Measure*". Proc. ICASSP 1993 Minneapolis, pp 375-378.

6. M. J. Carey and E. S. Parris. "*Cross Validation in Speaker Recognition*". Proc. RLA2C 1998 Avignon, pp 161-164.

7. A. Higgins et al. "*Speaker Verification Using Randomised Phrase Prompting*". Digital Signal Processing Vol. 1, 1991, pp 89-106.

8. M. J. Carey, E. S. Parris and J. S. Bridle. "*A Speaker Verification System Using Alpha-Nets*". Proc. ICASSP 1991 Toronto, pp 396-399.

9. A. Martin et al. "*The DET Curve in Assessment of Detection Task Performance*". Proc. Eurospeech 1997 Rhodes, pp 1895-1898.