Proceedings of the Institute of Acoustics

SPEAKER VERIFICATION.

Michael J. Carey, Eluned S. Parris and Stephen J. Bennett.

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

1. INTRODUCTION

This paper presents a summary of the state of the art in speaker recognition, giving an view of the main classes of techniques researchers are now investigating. Before proceeding with the paper a summary of terminology may be useful. We define as follows the main terms used in the field:

Speaker verification	The act of deciding whether an utterance from an unknown speaker was made by a specific speaker or not, an open set problem.
Speaker identification	The act of deciding which of a group of speakers made an utterance from an unknown speaker, a closed set problem.
Speaker Recognition	Either of the above.
Text Dependent System	A system in which the speaker is required to or is known to have uttered a known text.
Text Independent System	A system in which the speakers utterance is an unknown text.

For reasons of brevity we shall mainly discuss text independent speaker verification in this paper. Speaker identification systems are usually required to deal with the case where the unknown speaker is not one of the set equating the system to a multiple speaker verification system. Text dependent systems can be regarded as special cases of text independent systems where the statistics of the 'unknown' speech are predetermined. The main elements of a typical system are shown in Figure 1. A feature extractor is used to produce a set of features $O = O_1 \ldots O_t \ldots O_T$ which are applied to a pattern matching algorithm. The algorithm has access to speaker specific information derived from some training material which allows the building of an explicit or implicit speaker model, m_j. We can then estimate the probability of the observations given the model, $p(O \mid m_j)$, for each speaker of interest.

The pattern matching techniques used in speaker verification differ in the way this likelihood is estimated. An important distinction exists between those techniques for which

$$p(O \mid m_j) = \prod_{t=1}^{T} p(O_t \mid m_j) \quad \text{or in the log domain} \quad p(O \mid m_j) = \sum_{t=1}^{T} p(O_t \mid m_j) \qquad (1)$$

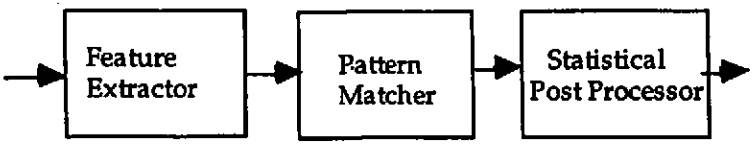

Figure 1. The Verification System

that is the likelihood of the whole sequence is simply the product of the likelihood of the individual frames, and those for which a more complicated formulation is used. In the former case, for example the vector quantiser the ordering of the sequence is immaterial. However for a Hidden Markov Model (HMM) system this is not the case as the temporal ordering of the frames affect the estimate of $p(O|m_j)$. In the remainder of this paper we first review techniques for feature extraction and then examine a number of pattern matching methods.

2. FEATURE SELECTION

2.1 Spectral Estimation.
Methods of spectral estimation closely follow those used in speech recognition. An estimate of the spectrum of the speech is produced at a frame rate of 10-20ms by Linear Prediction analysis, a fft or filterbank and these are orthogonalized by a discrete cosine transform to give linear or mel scale cepstral coefficients. Some authors have found that perceptually based linear prediction can be used to advantage[1].

2.2 Transitional Features.
As in speech recognition linear and quadratic estimates of the trend of the spectrum represented as estimates of the first and second derivatives of the cepstra are sometimes included in the feature vector[2]. The exact benefit of including these features particularly the second derivatives has not been firmly established. Our results indicate that when the parameters of the model corresponding to these features are well estimated they can increase the performance of the system. However with small amounts of training data this is sometimes not easy to do.

2.3 Channel Normalisation and Feature Transformations.
For systems working through variable communications channels such as the public switched telephone network a method of correcting for handset and channel variations

is required. This is frequently achieved by cepstral mean subtraction (CMS), that is estimating the means of the features, cepstra, over each segment of speech and subtracting the mean from the instantaneous estimate of the features hence removing the effect of the static channel[3].

Many of the channel variabilities arise from the differing characteristics of telephone handsets. Where these characteristics are linear and time invariant CMS can correct for them. However the continued use of carbon microphones particularly in the USA leads to non-linear and time variant distortion[4]. To date it has not been found possible to correct for this during feature analysis.

Linear Disciminant Analysis (LDA)[5,6] has been shown to be a worthwhile means of increasing the accuracy of speech recognisers. In that case the states or mixture components of the word models are treated as separate classes with intraclass variability being caused in part by inter-speaker differences. Unfortunately in speaker verification it is these intraclass differences which are important. While a LDA transform trained for speech recognition can be used for speaker verification by omitting the lowest, most significant dimensions for speech recognition, little advantage appears to result. Other methods of estimating transformations for speaker recognition have not been widely researched.

2. 4 Pitch

The use of the pitch of a speaker's voice for discrimination between speakers was the subject of research in the 1970's[7,8]. Research was subsequently directed towards the use of cepstral methods. Interest in its use has recently been growing, but as an adjunct to other techniques as it provides additional speaker discriminative information[]. Previously the pitch contour estimated frame by frame had been used as a feature vector. Now however more robust techniques such as the mean and variance of the pitch estimated over an utterance have been used as features [9].

3. VECTOR QUANTISERS

In Vector Quantiser systems each speaker is modelled by a codebook of vectors[10]. The codebook is trained using the Linde, Buzo and Gray algorithm[11] which attempts to minimise the global distances between the training vectors and their nearest codebook entry. The distance between feature vectors is defined as:

$$d(X,Y)=\sum_{k=1}^{K}(X_k-Y_k)^2 \quad \text{.....(2)}$$

In verification the nearest codebook entry, $a_{k,j}$, to each input frame is found and the distance from the input to that entry is accumulated into the overall score,

$$p(O_T \mid m_j)=\sum_{t=1}^{T}\min_{k}(d(a_{k,j},O_t)) \quad \text{.....(3)}$$

The essence of the technique is to use the LBG algorithm to find a set of codebook entries which model the speakers training vectors in the hyper-volumes of the features space in which they occur.

4. NEURAL NETWORKS

At least three types of neural net techniques have been researched, the Multi-layer Perceptron(MLP)[12], Radial Basis Functions(RBF)[13] and more recently the Neural Tree Network[14]. Time and space do not permit a detailed description of these techniques for which the interested reader is referred to the references. However the RBF has an interesting relation to the Vector Quantiser and the Gaussian Mixture Model. The RBF comprises a set of hidden units which are typically gaussian. The output of the network is the linearly weighted sum of the outputs of these units. Hence we have

$$p(O_\tau \mid m_j) = \sum_{t=1}^{T} \sum_{i} w_{i,j} \exp(d(a_{k,j}, O_t)) - T_j \quad \text{.....(4)}$$

where T_j is a bias term and w_{kj} are the network weights. The RBF replaces the min. in the vector quantiser with a weighted sum of gaussians. The network is initialised by finding the centroids of the basis functions by clustering as in the vector quantiser. Training can then be carried out by using gradient descent to find values of the weights which optimally discriminate between the target speaker and impostors.

5. GAUSSIAN MIXTURE MODELS

Gaussian Mixture Models(GMMs)[15] can be regarded as single state HMMs with a large number of mixture components, 128 or 256 are common and a unity self transition probability. Each speaker is modelled by a multivariate gaussian density in the feature space. It is universally acknowledged that the off diagonal components of the cepstral covariance matrix must be close to zero. Hence a diagonal covariance matrix is assumed, and the frame probability is given by

$$p(O_t \mid m_j) = \sum_{k=1}^{I} \frac{w_{k,j}}{2\pi \det \Sigma} \exp \sum_{k=1}^{K} -\frac{1}{2\sigma_{k,j,j}^2} (O_{tk} - a_{k,j,j})^T (O_{tk} - a_{k,j,j}) \quad \text{.....(5)}$$

The relation with the vector quantiser and the RBF network is clear since again we have weighted sums of gaussians. The GMM however has variances which are not necessarily unity and the training of the system takes a different form. Maximum likelihood training using some form of the Expectation Maximisation algorithm is used. The Baum-Welch algorithm is a popular choice. Good results have also been achieved with maximum a priori (MAP) estimation and unsupervised learning. The training unlike that of the RBF network is not discriminative.

6. HIDDEN MARKOV MODELS

6.1 Model Topologies

Hidden Markov Model systems have been used extensively in text dependent systems where word or phrase models can be constructed from training data comprising the known text[15,16]. In text independent systems[17,18,19] researchers have used subword models based on phone sized units. The subword units used can either be based on the full set of phonemes of the language or on a reduced set for example broad classes. Typically three state left to right models with no skips are used. The systems use continuous gaussian densities with diagonal covariances. The chief difference between these systems and the GMM systems is that the HMM imposes temporal constraints on the succession of densities matches both during testing and training.

6.2 Parameter Estimation

Maximum likelihood training using some form of the Expectation Maximisation algorithm is used. The Baum-Welch algorithm is the usual method. Good results have also been achieved with maximum a priori (MAP) estimation.

6.3 Pattern Matching

The pattern matching technique can take one of several forms. Generally the Viterbi algorithm is used to find the sequence of models which best explains the test utterance. The speakers score, $p(O \mid m_j)$, is then the likelihood along this path. While this can work well with utterances from the true speaker, impostor utterances can some times score well by matching improbable sequences of models. This problem has been addressed by pre-aligning the test utterance to a set of speaker independent models and then matching the corresponding speaker dependent models to sections of the test utterance matched to the speaker independent model. Alternatively the independent and dependent models can be combined into a single set allowing a combination of the two to be matched to the test utterance. The speaker score is then the proportion of the total models matched.

7 DECISION TECHNIQUES

7.1 Normalisation.

Normalisation is now considered a key requirement in speaker verification systems. The reader may recall from the introduction that we have estimated $p(O \mid m_j)$ but we should use $p(m_j \mid O)$ the probability or likelihood of the model given the observations. These are related by Bayes theorem,

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

Usually the prior probability of the speaker $p(m_i)$ is assumed to be the same for all speakers and is disregarded. Also $p(O) = \sum_i p(O \mid m_i)$,. Hence we have

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

Now $\sum_i p(O \mid m_i)$ is the sum of the likelihoods for all possible speakers, a normalisation of $p(O|mj)$. The exact evaluation of $p(O)$ is clearly impossible Therefore two approximations to this have been proposed. The first relies on the observation that $p(O|mj)$ will be small for all but a small set of similar speakers and that the approximation $\sum_i p(O \mid m_i) \approx \sum_{i \in A} p(O \mid m_i)$ 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[16]. The other approach is to construct a 'world' or 'general' model M which may for example be speaker independent model (M) in the case of a Hidden Markov Model system[15]. We then have $p(m_j \mid O) = \frac{p(O \mid m_j)}{p(O \mid M)}$, 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[20].

7.2 Handset Modelling

While Cepstral Mean Normalisation reduces the linear effects of handset variability problems arise when carbon microphones are used since they produce non-linear distortion of the speech waveform. This produces intermodulation components in the spectrum which manifest themselves as additional spectral components particularly observable at low levels. Although telephones with carbon microphones are now a rarity in the British public switched telephone network this is not the case in the USA. Also a similar type of distortion can be introduced by speech coding algorithms such as the ones used in digital mobile telephones. While the solution to this problem may be a feature set which is unaffected by the presence of this form of no such feature set has been found to date. The present approach is to detect the mismatch between microphone types in training and testing and then to vary the value of the normalisation constant from the default value when a mismatch between training and testing handset is detected.

8. PERFORMANCE

In trying to describe the performance of speaker verification systems one is confronted by the problem of test databases. The performance of the algorithms vary substantially between databases making comparative statements difficult and absolute conclusions impossible. Nevertheless this is what we shall try to do.

Surprisingly perhaps given the high hopes workers had for MLPs they give the worst performance of all the techniques discussed. Both VQs and RBFs give superior

performance[21,22]. VQs has similar performance to NTNs[21] but are poorer than RBFs[13]. However GMMs perform better than VQs provided enough training data is available[22].

GMMs have also performed better in tests than HMMs as the results of the NIST evaluations show[23]. The differences between the GMMs and the HMMs are twofold. The GMM uses unsupervised learning and the Hidden Markov Model has temporal constraints. Tishby[24] however demonstrated that there is speaker specific information in the state durations and transitions of the HMM. We have carried out experiments in which distributions estimated for HMMs using supervised training are using in a GMM and a HMM. The HMM then has better performance than the GMM showing that the retention of the durational information improves the results. On the same data a GMM using distributions estimated for by unsupervised training performed better than the HMM. We hypothesise that since the GMM can pool data from different phonemes to estimate the parameters of a single distribution better estimates of these parameters are made leading to the superior performance already noted. This and other aspects of speaker verification require further investigation.

The performance of the best systems on telephone quality speech can best be demonstrated by quoting the results from the best systems submitted for the NIST 1996 evaluation. With two minutes of training speech and 30s test files from the training handset the systems achieved an equal error rate of 3%. This increases to 5% for 10s testing files and 7% for 3s test. The performance was somewhat worse when the test files were taken from other handsets. It is interesting to note that the error rate in these tests has halved between the 1995 and 1996 evaluations.

9. REFERENCES

[1]H. Hermansky, 'Perceptual Linear Predictive(PLP) Analysis for Speech', J. Acoustic of America 1990, pp1738-1752.

[2]S. Furui, 'Cepstral Analysis Techniques for Automatic Speaker Recognition', IEEE Trans. ASSP, Vol. 29, April 1981, pp254-272.

[3] T. Stockham, T. Cannon, and R. Ingebretsen, 'Blind Deconvolution through Digital Signal Processing', Proc. IEEE, Vol63, April 1975, pp 678-692.

[4]L. Moye, 'Study of the Effects on Speech Analysis of the Types of Degradation Occurring in Telephony', STL Monograph No1 July 1979

[5]M. Hunt et al. "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination", Proc ICASSP 91.

[6]E. Parris and M. Carey, "Estimating Linear Discriminant Parameters for Continuous Density Hidden Markov Models", Proc ICSLP 94

[7] B S Atal, 'Automatic Speaker Recognition Based on Pitch Contours', JASA Vol.52 No.6 pp 1687-1697.

[8] J D Markel, B T Oshika and A H Gray, 'Long-Term Feature Averaging for Speaker Recognition', IEEE Trans ASSP Vol. ASSP-25, pp 330-337.
[9] M. Carey, E. Parris, S. Bennett, and H Lloyd-Thomas, 'Robust Prosodic Features For Speaker Identification' Proc. ICSLP 1996.
[10] F Soong et al., 'A Vector Quantisation Approach to Speaker Recognition', ATT Tech J. Vol. 66, pp14-26.
[11] Y. Linde, A. Buzo and R. Gray, 'An Algorithm for Vector Quantisation', IEEE Trans Comms Vol. COM-28, pp84-95.
[12] J. Oglesby and J. Mason, 'Optimisation of Neural Models for Speaker Identification. Proc. IICASSP1990, pp261-264.
[13] J. Oglesby and J. Mason, 'Radial Basis Function Networks for Speaker Recognition', Proc. ICASSP1991 pp393-396
[14] HS. Liou and R. Mammone, 'A Subword Neural Tree Network Approach to Text-Independent Speaker Verification', Proc. ICASSP 1995, pp357-360.
[15] M. J. Carey, E. S. Parris and J. S. Bridle 'A Speaker Verification System Using Alphanets', Proc ICASSP 1991 pp397-400.
[16]A. Higgins et al. 'Speaker Verification Using Randomised Phrase Prompting', Digital Signal Processing Vol.1. 1991, pp89-106.
[17] R. Rose and R Reynolds, 'Text Independent Speaker Identification Using Automatic Acoustic Segmentation', Proc ICASSP 1990 pp293-296.
[18] A Rosenberg et al. 'Sub-word Unit Talker Verification Using Hidden Markov Models', Proc. ICASSP 1990 pp269-272.
[19] E. Parris and M. Carey, "Discriminative Phonemes for Speaker Identification", Proc ICSLP1994 pp1815-1818.
[20]A. Setlur and T Jacobs, 'Results of a Speaker Verification Service Trial Using HMM Models', Proc Eurospeech 1995 pp639-642.
[21]K. Farrell et al, 'Speaker Recognition Using Neural Networks and Conventional Classifiers', IEEE Trans. On Speech and Audio Processing, Vol. 2 No.1, January 1994, pp194-205.
[22] Kim Yu et al., 'Speaker Recognition Models', Proc Eurospeech 1995 pp 629-632.
[23]NIST, Proc. March 1996 Speaker Identification Workshop.
[24]N. Tishby, 'On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition', IEEE Trans. on Signal Processing, Vol. 39 No. 3, March 1991, pp563-570.