A COMPARISON OF MODEL ESTIMATION TECHNIQUES FOR SPEAKER VERIFICATION

Michael J Carey, Eluned S Parris, Stephen J. Bennett and Harvey Lloyd-Thomas.

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

ABSTRACT

In this paper we address the problem of building speaker dependent Hidden Markov Models for a speaker verification system. A number of model building techniques are described and the comparative performance of a system using models built using each of these techniques is presented. Mean estimated models, models where the means of the HMMs are estimated using segmental K means but where the variances are taken from speaker independent models, out performed other techniques such as Baum Welch re-estimation for training times of 120s, 60s and 15s. Mean estimated models were also built with varying numbers of components in the state mixture distributions and a performance gain was again observed. The incorporation of transitional features into the system had degraded performance when the Baum-Welch algorithm was used for model estimation. However the inclusion of delta and delta-delta cepstra into the system using mean estimated models now gave a significant improvement in performance. Taken together these changes halved the equal error rate of the system from 15.7% to 7.8%.

1. INTRODUCTION

Enrolment material for training speaker recognition systems is often limited. The algorithm which estimates the parameters of the speaker's models must be able to capture the essential features of a speaker's voice from the training material if the speaker verification system is to perform well. Specifically the estimation program should produce accurate estimates from the minimum training data. This parsimonious use of the training data allows for the possibility of:

- Using more complex models to better represent the underlying distribution.
- Setting speaker specific thresholds by reserving some of the training data for testing.
- Building at least some of the set of models when data is scarce.

Therefore we investigated several techniques for the estimation of the parameters of the target models, these are described in Section 2. We then use the best technique to estimate models with multiple mixture components and enlarged feature sets.

2. MODEL ESTIMATION METHODS

Six different techniques were used in the experiments described. The techniques can be summarised as follows:

1. Expectation Maximisation - using the Forward Backward Algorithm, Baum Welch Re-estimation. This was regarded as the reference technique since it is widely used and we had used this previously (BW)[1].
2. Segmental K Means - frames in the training data are aligned to the states of speaker independent models using the Viterbi algorithm and the speaker dependent state means and variances are computed from the frames aligned to the same state. (SKM)[2].
3. Maximum A Priori - estimates of the means and variances of the target models by modifying a speaker independent model using the parameters of a SKM speaker dependent model built on the training data.(MAP)[3].
4. Maximum A Priori Means - estimates of the means only of the target models by modifying a speaker independent model using the parameters of a SKM speaker dependent model built on the training data. The rate of adaptation is dependent on the variance of the speaker independent model (MAP-Mean)[3].
5. Fixed Weighted Average - as MAP Means but the mean estimate is adapted by some predetermined constant independent of the variance. (FWA)
6. Mean Estimation - the means are estimated as in the SKM but the variances are fixed as in Maximum A Priori Means. This can also be regarded as Fixed Weighted Average with a very high weighting towards the speaker dependent models. (ME)

3. NORMALISATION.

Normalisation is considered a key requirement in speaker verification systems. We note that the pattern matching stage of the system estimates $p(O \mid m_j)$ the likelihood of the observation sequence given the model but we require $p(m_j \mid O)$ the likelihood of the model given the observations. These are related by Bayes theorem,

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

The prior probability of the speaker $p(m_j)$ is usually assumed to be the same for all speakers and is disregarded. Also in this

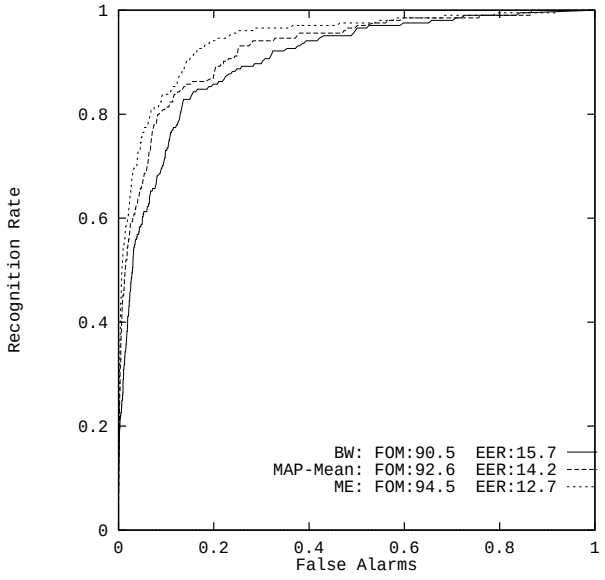

Figure 1: ROC curves showing relative performance of BW, MAP-Means, and ME models at 120s. One- session, 10s test.

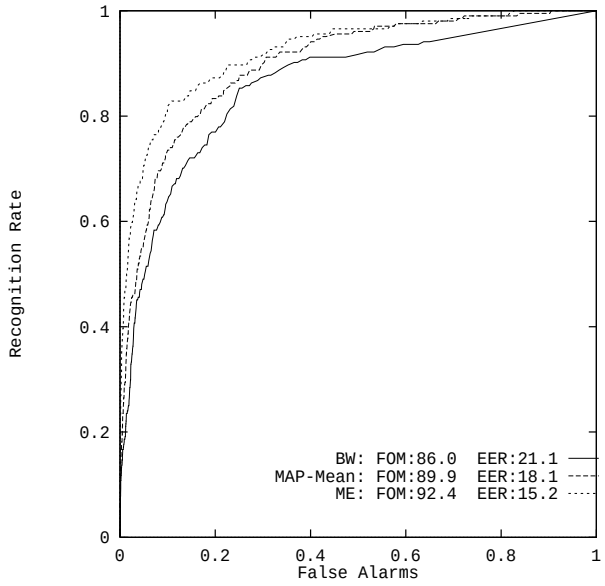

Figure 2: ROC curves showing relative performance of BW, MAP-Means, and ME models at 30s. One- session, 10s test

case we have $p(O) = \sum_i p(O \mid m_i)$, where m_i are the models for other speakers. 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. The exact evaluation of $p(O)$ is clearly impossible. Therefore two approximations to this have been proposed. The first relies on the observation that $p(O \mid m_j)$ 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[4]. The other approach is to construct a 'world' or 'general' model M which may for example be a speaker independent model (*M*) in the case of a Hidden Markov Model system[5]. We then have $p(m_j \mid O) = \frac{p(O \mid m_j)}{p(O \mid M)}$, that is the speaker independent 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[6]. This form of normalisation was therefore used in the experiments reported here.

4. EXPERIMENTAL CONDITIONS

4.1 System Description.

There follows a brief description of our system, for a fuller system description the speaker is referred to [7] and [8]. 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 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 segment of speech and subtracted from each of the feature vectors. Twenty eight subword models were used to model the phones of each target speaker. The stops and fricatives were each combined into a single broad-class model while each of the other phones was represented by its own model. The subword models used were three state Hidden Markov Models with continuous mixture distributions and a left to right topology and no skipping of states allowed. State distributions were characterised by a diagonal covariance matrix.

Speaker independent models were built using the TIMIT database and the American-English part of the OGI Multilingual Corpus. Each model state had seven Gaussian mixture modes. At training time these speaker independent models were used to segment the training speech for each of the test speakers and speaker dependent models were then built from this speech. Initially each of the speaker dependent models had a single mode per state. This was later increased as explained in Section 5 below.

During recognition an unknown speaker's speech was matched to a set of models comprising each of the hypothesised target speaker's dependent models and a set of speaker independent models trained on Switchboard data from the NIST 1995 evaluation. A score was generated for each of the target speakers which was the percentage of the total matches achieved by that speaker's models.

4.2 Test Corpus.

The experiments described in this paper were carried out on the development data for the NIST 1996 speaker identification

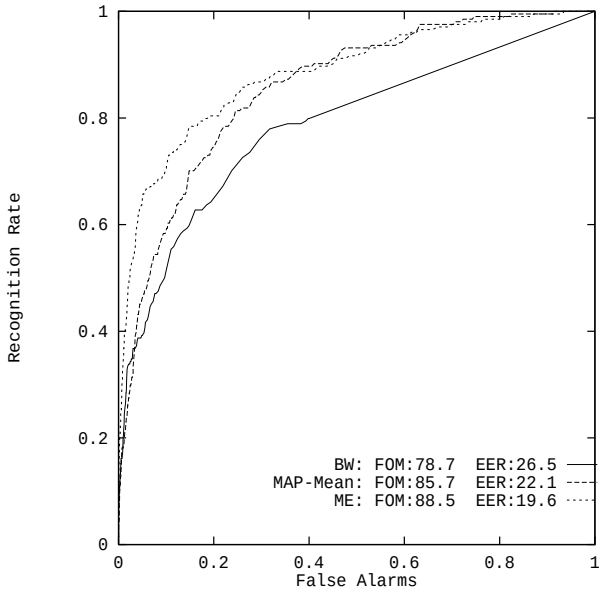

Figure 3: ROC curves showing relative performance of BW, MAP-Means, and ME models at 15s One-session 10s test.

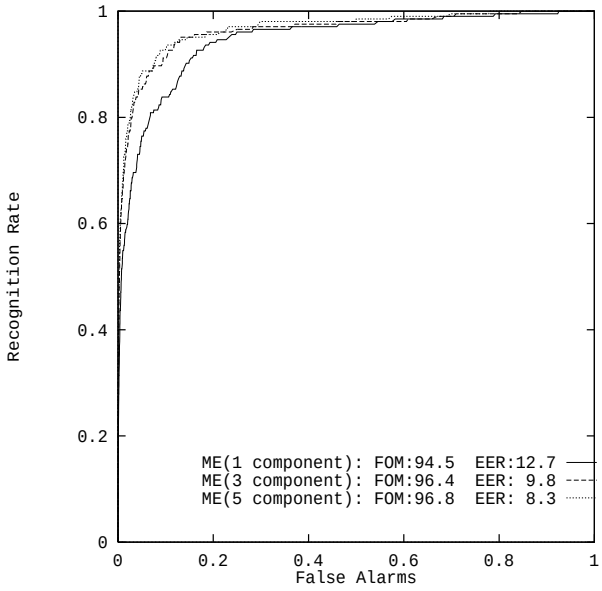

Figure 4: ROC curves showing relative performance of System with varying numbers of mixture components in the models, Means Estimated models with 120s Training. One-session, 10s test.

evaluation. This comprised 43 male and 45 female target speakers for whom four minutes of training data were provided. Training conditions consisted of one minute from a conversation with a second minute from either

1. the same conversation.
2. a different conversation using the same handset, actually the same telephone number.
3. a different conversation using a different handset.

The test material consisted of five files from each speaker from which either 3s, 10s, or 30s of speech was used. The task was to score each test file for each of the speakers against the models for all the speakers and then generate a Receiver Operating Characteristic curve from the scores. The experiments we carried out on this data used training material from the same conversation and the 10s set of test data.

5. RESULTS

5.1 Model Building Algorithms

Initial experiments were carried out using the 120s training data and the male speakers only. The results shown in Table 1 indicate that of the two full re-estimation techniques BW outperforms SKM, of the two MAP techniques MAP-Means outperforms MAP and of the two other techniques that ME outperforms FWA. Results from other experiments, for example using female speakers, confirmed this hence the further results presented are restricted to these three techniques.

The time available for training was then reduced. When for Baum-Welch re-estimation there were no examples in the training data to train a speaker dependent model the equivalent independent model was substituted. At the shortest training times there were sometimes no training examples of particular models. When this occurred the model set was completed by including the equivalent speaker independent model although this was excluded from the subsequent scoring. Figures 1 to 3 show the effect of reducing the training time on the ROC curves. These results demonstrate that BW re-estimation requires more data than the other two methods. This is unsurprising since unlike the other methods estimates of the parameter variances are made by the BW algorithm. It is well known that more data is required to make an estimate of the same accuracy of the variance of a distribution than the estimate of the mean. Of more interest is that the ME method outperforms the MAP-means technique. The inclusion of speaker independent data into the estimation process does not appear to give more robust models. There is also only a small degradation in the performance of the ME models when the training time is reduced from 120s to 30s. Hence using these models some of the training material could be retained for use in calibration test scoring allowing speaker dependent thresholds to be set.

	BW	SKM	MAP	MAP-M	FWA	ME
FOM	90.5	89.9	87.8	92.6	94.4	94.5
EER	15.7	17.2	20.6	14.2	13.2	12.7

Table 1. Performance of the Six Model Building Techniques, FOM is figure of merit while EER is equal error rate.

5.2 Multiple Mixtures

Our next experiment was to try and improve the performance of the system with the full two minutes of training data by increasing the number of components in the state mixture distributions from one to five. Previous experiment had shown that BW models with three component mixtures performed less well than those with a single component when trained with two minutes of speech. However since we were able to estimate

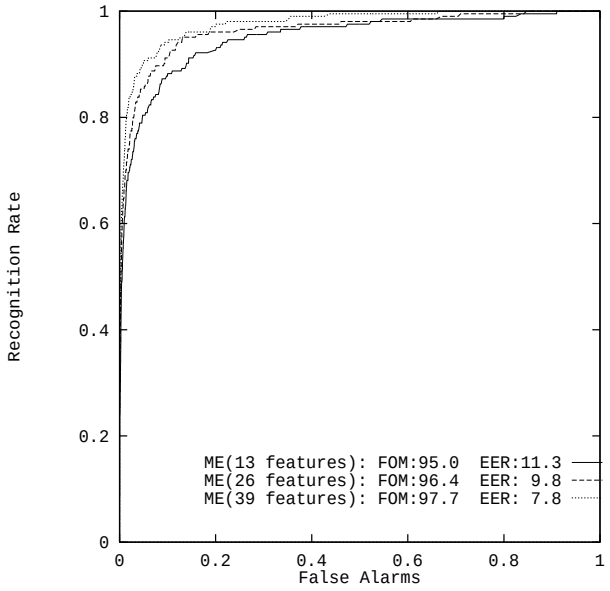

Figure 5: ROC curves showing relative performance of System with and without Delta and Delta-delta cepstra, Means Estimated models with 120s Training. One-session, 10s test.

good ME models with less training than BW required we modified the ME technique to produce three and five component models. The training data was segmented by aligning the speech with the speaker independent models using the Viterbi algorithm. Frames which were aligned to the same mixture component along the optimal path were pooled and the means of the distributions were estimated as the means of the pool. The variances of the distributions were set equal to the corresponding component variances in the speaker independent model. Figure 4. shows the ROC curve for these models and demonstrates that a significant improvement has been achieved.

5.3 Delta -Delta Cepstra

The second derivative is widely used in speech recognition and has been shown to give improved performance in a speaker verification task[9]. When added to our system using the BW for model estimation the performance of the system was degraded. However when the ME technique was used the results are as illustrated in Figure 5 which shows that adding first delta cepstra and then the delta-delta cepstra produce incremental improvements. We surmise that the variances were not estimated correctly when Baum-Welch re-estimation was used causing a degradation instead of an improvement.

6. CONCLUSIONS

In this work we have examined a number of approaches to the problem of building models for speaker verification. We have shown that the best performance for all training times is given by models in which only the means of the distributions are estimated from the training data and the variances are set the same as those of the distributions of the corresponding speaker independent models. This also gives good estimates of the parameters of the derivatives of the cepstra when these are included in the feature vector. Also when sufficient training data is available increasing the number of mixture components in the states of the subword model improves performance.

REFERENCES

[1] J. Holmes, “Speech Synthesis and Recognition” Chapman and Hall 1988 ,pp 137-145.

[2] L. Rabiner and B. H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall 1993, pp 382-383.

[3] C. H. Lee et al, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models”, IEEE Trans on Signal Processing Vol. 39. No.4 April 1991, pp 806-812.

[4] A. Higgins et al. “Speaker Verification Using Randomised Phrase Prompting”, Digital Signal Processing Vol.1. 1991, pp89-106.

[5] M J Carey, E S Parris and J S Bridle, “A Speaker Verification System Using Alpha-Nets”, Proc. ICASSP 1991, Toronto pp pp397-400

[6] [] A.Rosenburg and S. Pathasarathy, “Speaker Background Models for Connected Digit Password Speaker Verification”, Proc ICASSP 1996 Atlanta, pp 81-84

[7] E. S. Parris and M. J. Carey, “Discriminative Phonemes for Speaker Identification”, Proc ICSLP 94, Yokohama pp 1843-1846

[8] M. J. Carey, E. S. Parris, H. Lloyd-Thomas and S. J. Bennett “Robust Prosodic Features For Speaker Identification”, Proc ICSLP 96 pp1800-1803.

[9] S. Furui, “Cepstral Analysis Techniques for Automatic Speaker Recognition”, IEEE Trans. ASSP, Vol. 29, April 1981, pp 254-272.