IMPROVING A GMM SPEAKER VERIFICATION SYSTEM BY PHONETIC WEIGHTING

Roland Auckenthaler, Eluned S. Parris and Michael J. Carey
Ensigma Ltd., Turing House, Station Road, Chepstow, Monmouthshire, U.K.
{roland eluned michael}@ensigma.com

ABSTRACT

This paper compares two approaches to speaker verification, Gaussian mixture models (GMMs) and Hidden Markov models (HMMs). The GMM based system outperformed the HMM system, this was mainly due to the ability of the GMM to make better use of the training data. The best scoring GMM frames were strongly correlated with particular phonemes e.g. vowels and nasals. Two techniques were used to try and exploit the different amounts of discrimination provided by the phonemes to improve the performance of the GMM based system. Applying linear weighting to the phonemes showed that less than half of the phonemes were contributing to the overall system performance. Using an MLP to weight the phonemes provided a significant improvement in performance for male speakers but no improvement has yet been achieved for women.

1. INTRODUCTION

The two most common and successful approaches to text independent speaker verification are based on modelling the speech by Gaussian mixture models (GMMs) [1] and hidden Markov models (HMMs) [2]. In a HMM based system, speaker verification is usually performed using a phonetic description of the incoming speech. Temporal information is modelled by the HMMs and consecutive frames of data are forced to align to a sequence of sounds of the language being spoken. A GMM based system uses a general description of the data and each frame of features generated is treated independently of all other frames.

In recent speaker recognition evaluations carried out by the National Institute of Standards and Technology (NIST), the best GMM based systems have outperformed the HMM based systems [3]. This suggests that no gain in performance is being achieved by the use of temporal information captured in the HMMs. The research described in this paper has concentrated on investigating the difference between the two approaches. Experiments have been carried out to try and improve the performance of a GMM system by using phonetic knowledge contained in the HMM system.

Section 2 describes the experimental configuration of the GMM and HMM systems and provides a description of the databases used. Section 3 compares the results from a series of experiments using two speaker verification systems with identical front end processing, one based on GMMs and the other on HMMs. The interaction between GMMs and the phonetic labelling produced by a HMM system is described in Section 4. Section 5 presents the improvements made to a GMM based speaker verification system by using phonetic weighting and Section 6 presents the conclusions of this work.

2. EXPERIMENTAL CONFIGURATION

The experiments described in this paper were carried out using the NIST 1998 Evaluation data. The acoustic analysis used in the experiments was as follows. The data was sampled at 8kHz and then filtered using a filterbank containing nineteen filters. The log power outputs of the filterbank were transformed every 10ms into twelve mel frequency cepstral coefficients (mfcc) and their first and second derivatives. Further the frame energy, delta energy and delta-delta energy were used to represent the input speech by thirty-ninth order feature vectors. The mean of each of the cepstral parameters was estimated for each segment of speech and subtracted from each of the feature vectors.

For the comparison of the two systems, GMM and HMM, identical front ends were used. Speaker independent background models were created for both systems using eight hours of speech taken from the OGI Multilingual Corpus [4] and from the NIST 1995, 1996 and 1997 speaker recognition evaluations. The data was equally distributed across each gender and database.

The background models for both systems were trained using the EM-algorithm [5]. The GMM background model consisted of 256 mixtures and was built from data of both genders. Two sets of 28 HMMs based on phonetic classes were built, one for each gender, in addition to general models representing noise. Each HMM had three states with three Gaussian mixture modes per state. A left to right topology was used with no skipping of states allowed. The number of Gaussian distributions used were chosen to minimise the differences between the two systems. The GMM system had a total of 256 Gaussian distributions compared with 252 for the HMM system.

For both systems, models were trained for each target speaker using mean only adaptation. The variances of the distributions were taken from the associated modes of the speaker independent background models.

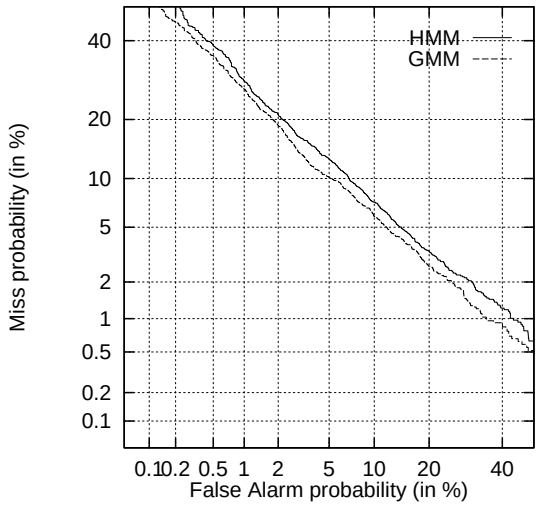

Figure 1: Speaker Recognition Results for GMM and HMM Systems, One-Session Train, 30s Test.

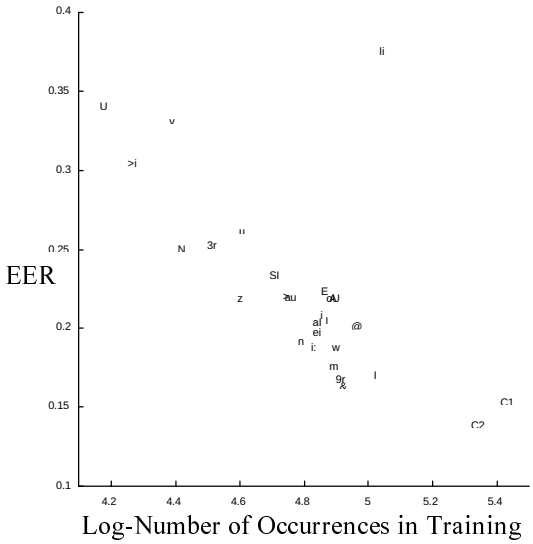

Figure 2: Speaker Recognition Error vs. Number of Occurrences Seen in Training

In the HMM system, scoring was performed by accumulating the log likelihood ratios of the 65% best scoring vectors. In the GMM system, a general noise model was used to reject non-speech frames and all of the remaining frames were used to calculate the speaker score. These methods were optimised on previous speaker recognition evaluation data.

The speaker scores were normalised for both systems by the background model [6] and by z-normalisation [7] using 100 speech files for each handset (electret and carbon) taken from the NIST 1997 Evaluation.

3. GMM HMM SYSTEM COMPARISON

The systems described above were tested using the 30s test conditions of the NIST 1998 Evaluation. Both genders were used to produce the results. A DET-plot [8] for both systems is shown in Figure 1 for the one-session training condition. The GMM system performed significantly better than the HMM system for each of the training conditions.

It is believed that the main difference in the performance was due to the training of the models used. In the HMM system, the training speech is labelled with a phonetic based transcription and the phoneme specific frames are uniquely assigned to one of the HMM phoneme models. In the GMM system, one large model is used allowing the sharing of training data between different mixtures, disregarding the phoneme specific information. This leads to better trained mixture parameters. Therefore it would first appear that the GMM does not use any phonetic knowledge of the incoming speech. The HMM is able to use this knowledge along with temporal modelling but does not perform as well as the GMM based system because of poorly trained distribution parameters.

4. GMMS AND PHONETIC CLASSES

The results described above led to further investigation of the best scoring GMM frames. It was found that there was an exact correspondence between the best scoring frames of the GMMs and particular sounds occurring in the speech. The discrimination between speakers provided by the GMM was in fact due to a subset of phonetic classes. Even though the sounds were not being explicitly modelled by a GMM the scoring appeared to be closely related to them. Previous experiments with HMM systems have shown that the phonetic classes provide different amounts of discrimination between speakers [2,9]. Therefore these experiments were repeated using a GMM based system.

The positions of the phonemes in the test data were found by using the HMM based system to label the speech with a phonetic transcription. The frame likelihoods generated by the GMMs were then pooled for each phoneme separately and final scores obtained by normalising by the total number of frames in the file. Results for the different phonemes are given in Figure 3. This shows that certain phonemes perform significantly better than others and correspond closely to those providing the most discrimination in other systems [2,9].

Figure 2 shows the error rate for each phoneme compared to the number of frames of data seen at training time. While there is a strong correlation between performance and the amount of training data, there is also a marked difference in equal error rates between phonemes with the same amount of data. This shows that the amount of training data is not the sole explanation for the difference in verification performance of the phonemes. The GMM system is therefore also using discrimination based on the phonemes for speaker recognition.

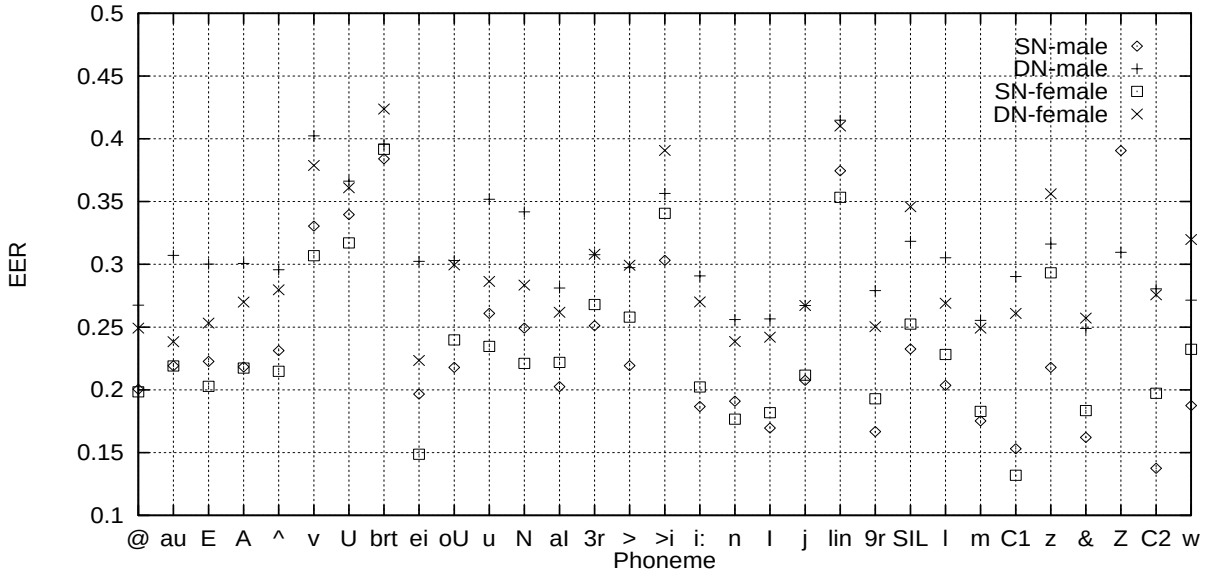

Figure 3: Speaker Recognition Performance For Each Phoneme Using A GMM System: Results show equal error rates (EER) for both genders and same number (SN) and different number (DN) conditions. C1 represents a collapsed phoneme class of ph, b, th, d, kh, g and h and C2 represents f, T, D, d(, s, S dZ and tS.

5. PHONETIC WEIGHTING

5.1 Linear Combination of Phoneme Scores

The GMM based system combines the information provided by the phonemes by adding the frame likelihoods to give a final score. This gives each of the phoneme classes the same weighting in the overall scoring.

The GMM system, whose performance is given in Figure 1, uses all of the phonemes present to produce the final score. However some of the phonemes provide little or no discrimination between speakers and the inclusion of these may lead to a degradation in performance. Therefore it should be possible to achieve a better system performance by omitting or reducing the contribution of these phonemes.

Figure 4 shows that using a subset of the most discriminative phonemes gives better performance than using all of the phoneme classes. Here the phonemes are sorted by their individual recognition performance and the scores of the n-best phonemes are summed with equal weight to produce the final score. The horizontal lines show the baseline performance achieved by the system using all of the phonemes for scoring for the same number and different number tests.

Figure 4 also shows that a small number of phonemes are responsible for the overall performance of the whole system. Using a system with a subset of 10 to 15 phonemes provides better results than the standard system using all of the phonemes. Although the difference in performance is not highly significant it may be possible to use more sophisticated techniques in the future to combine the phoneme scores.

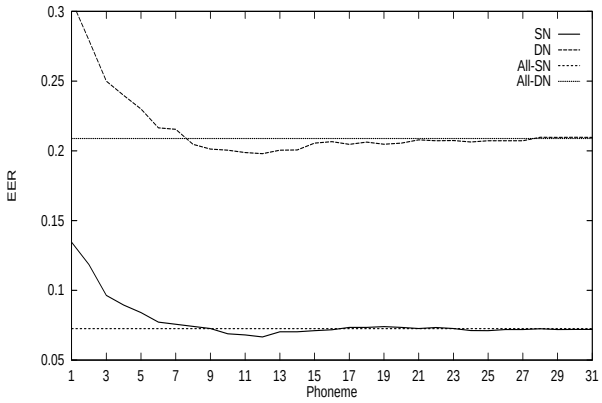

Figure 4: Speaker Recognition For Men Using A Subset Of All Phonemes.

5.2 Combining Phoneme Scores Using An MLP

As described previously, a subset of the 28 phoneme classes leads to the same result when the individual scores are added linearly. In general a linear summation does not lead to an optimal solution. Therefore an MLP was applied to the merging of the phoneme scores to try and improve performance.

For this purpose an MLP with one hidden layer containing fifty hidden nodes was used. The inputs to the net were the best fifteen phoneme classes, namely @, E, ei, aI, I:, n, l, j, 9r, I, m, C1, &, C2 and w. C1 and C2 are two collapsed consonant and fricative classes previously used in the HMM system. These phonemes had produced good results for both male and female speakers.

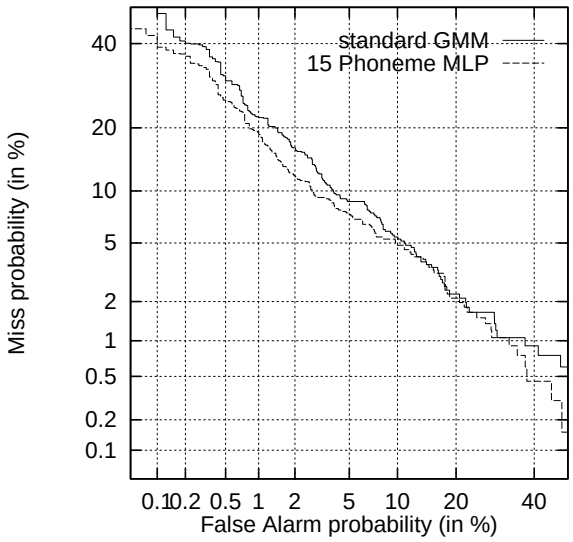

Figure 5: DET-Plot For The Male Speaker Set Using the Standard GMM And Fifteen Phoneme MLP.

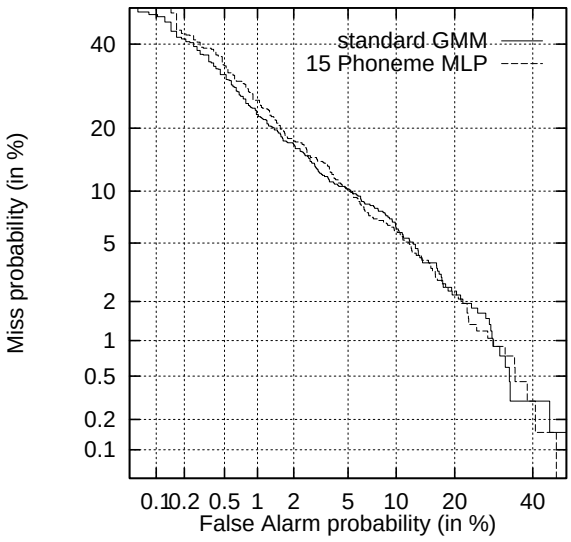

Figure 6: DET-Plot For The Female Speaker Set Using The Standard GMM And Fifteen Phoneme MLP

In training the MLP, not all of the phonemes scores were available for each speaker since the phonemes do not always occur in the test data. The missing scores were replaced by zero so that the training of the MLP was unaffected. It was discovered that the discrimination provided by the phonemes was different for men and women. Therefore, separate MLPs were then built for each gender.

The training and testing of the MLP shown here used the one-session 30 second test. The whole test for a particular gender contains 25000 individual tests with 250 target speakers and 2500 speech files. To provide independent train and test sets, the data was partitioned into three sets with each target speaker only occurring in one of the sets. The MLP was trained with one quarter of the files and a preliminary test performed using a different quarter of the files. A final test was then made using the remaining files.

Results using the MLPs are shown in Figures 5 and 6 for both genders. Results show that a significant improvement in performance has been made for male speakers by using the MLP approach. However no improvement has yet been achieved for female speakers.

6. CONCLUSIONS

This paper has provided a comparison between two text independent speaker verification systems with identical front end processing, one using GMMs and the other HMMs. The GMM based system provided significantly better performance for all tests. This was mainly due to the training of the GMM where the data is shared between the mixtures of one model. In training the HMMs, data is first aligned to phoneme classes and no data is then shared between classes when training the phoneme based HMMs.

The best scoring frames of the GMMs were highly correlated with particular phonemes. The performance was found to be mainly independent of the number of occurrences seen in training. Ten to fifteen phonemes were in fact contributing most to the discrimination between speakers. Applying both linear weighting to the phonemes and non-linear weighting by using an MLP, gave some improvement in system performance. This improvement was greatest for male speakers. Future work will concentrate on trying to incorporate more of the knowledge contained in the HMMs e.g. temporal modelling to improve GMM system performance.

7. REFERENCES

[1] D. A. Reynolds, 'Speaker Identification and Verification using Gaussian Mixture Speaker Models', Speech Communication, vol.17, August 1995, pp91-108.

[2] E. S. Parris and M. J. Carey, 'Discriminative Phonemes for Speaker Identification', Proc. ICSLP 1994, Yokohama, pp1843-1846.

[3] M. A. Przybocki and A. F. Martin, 'NIST Speaker Recognition Evaluation - 1997', Proc. RLA2C 1998, Avignon, pp120-123.

[4] Y.K. Muthusamy, R.A.Cole and B.T. Oshika, 'The OGI Multi-Language Telephone Speech Corpus', Center of Spoken Language and Understanding, OGI

[5] A. Dempster, N. Laid and D. Rubin, 'Maximum likelihood from incomplete data via the EM-Algorithm', J. Royal Stat.Soc, vol 39, 1977, pp38.

[6] M. J. Carey, E. S. Parris and J. S. Bridle, 'A Speaker Verification System using Alpha-Nets', Proc. ICASSP 1991, Toronto, pp397-400.

[7] D. A. Reynolds, 'Comparison of Background Normalisation methods for text-independent speaker verification', Proc. EUROSPEECH 1997, Rhodes, pp 963-966.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, 'The DET Curve in Assessment of Detection Task Performance', Proc. Eurospeech 1997, Rhodes, pp1895-1898.

[9] J. Eatock and J.Mason 'A Quantitative Assessment of the Relative Speaker Discriminative Properties of Phonemes', Proc. ICASSP 1994, Adelaide pp. I133-136.