Score Weighting in Speaker Verification Systems

Mohaddeseh Nosratighods[1,2], Eliathamby Ambikairajah[1,2], Julien Epps [2,1] and Michael Carey[3,1]
[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
[3]School of Engineering, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

m.nosratighods@student.unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au, m.carey@bham.ac.uk

Abstract— **This paper presents a method for re-weighting the frame-based scores of a speaker recognition system according to the discrimination level of the best matched Gaussian mixture for that frame. This approach focuses on particular feature space regions that either have been modeled accurately or contain the phonemes which are inherently most discriminative. The performance of individual Gaussian mixtures in terms of Equal Error Rate (EER) and minimum Detection Cost Function (DCF) on training, development and testing datasets consistently suggest that some Gaussian mixtures are inherently more discriminative regardless of their occurrence in training data. Therefore, it is possible to enhance the performance of speaker verification systems by re-weighting the frames that are mainly produced by those discriminative Gaussian mixtures. Compared with the baseline, results show a relative improvement of 5.82% and 5.46% on male speakers from the NIST 2002 dataset, in terms of EER and min DCF, respectively.**

Keywords—frame-based log-likelihood ratio, score modification, speaker verification

I. INTRODUCTION

Decision-making is the final processing stage of the speaker verification system, preceded by feature extraction and speaker modeling. The decision-making process employs a threshold, with which the log likelihood ratio (LLR) resulting from the claimed speaker model and the general population model for a given test utterance are compared [1]. A problem arises when the matching score of a claimant model varies across frames [2]. The rate of change of score distributions suggests that it is closely related to the phonetic content of the unknown speech, i.e. some segments of unknown speech are poorly matched due to the lack of training data for those particular phonemes. In addition to the lack of training data for particular phonemes, a number of studies have indicated that not all phonemes are equally discriminative [3]. Even though there is a strong correlation between the performance of the system and the amount of training data for a particular phoneme, the error rates are not equal even with the same amount of training data for any given phonemes, as the distributions will be different [3]. The lack of available training data in speaker verification was the motivation for using MAP adaptation [4], [5] to model the characteristics of a specific speaker. However, the assumption that the universal background model is representative of the acoustic regions of the feature space that are not accurately updated, due to a lack of training data, is not always valid. Thus phonetic content variation, in addition to other factors such as the variability of the feature vector distribution from session to session, and MAP adaptation itself, has made some regions of the feature space less reliable in making the final decision. We have addressed the score variability caused by the lack of training data in our previous work [6,7] by dropping the non-discriminative frames according to their target and impostor scores without making any *a priori* assumptions about the distributions of impostor and target scores.

Following on from our previous investigations [6,7], we now address the score variability caused by phonetic variation by emphasising the best scoring GMM frames that are strongly correlated with particular phonemes e.g. vowels and nasals [3]. An information measure [8] for each Gaussian mixture is introduced which is based on the occurrence of particular Gaussian mixtures in the available training data compared with the general public model. The performance of speaker verification systems for any individual Gaussian mixture is consistent on the training, development and testing data in terms of EER and min DCF. It has been shown that even if the most informative Gaussian mixtures generally result in better performance, this is not always the case. Therefore, the performance of any speaker verification system also depends on the varying amounts of discrimination provided by the phonemes [3]. An accurate phone recognizer is required to label the speech signals with phonemes however Gaussian mixture components can be considered to be representative of the phonemes in these features space.

Section 2 explains the need for re-weighting the frame-based scores and the relationship between the discrimination level of different regions of feature space and the Gaussian mixtures which model them. Then we describe the system setup (Section 3), and experiments selecting a sub-set of the most discriminative Gaussian mixtures for the final decision are reported in Section 4.

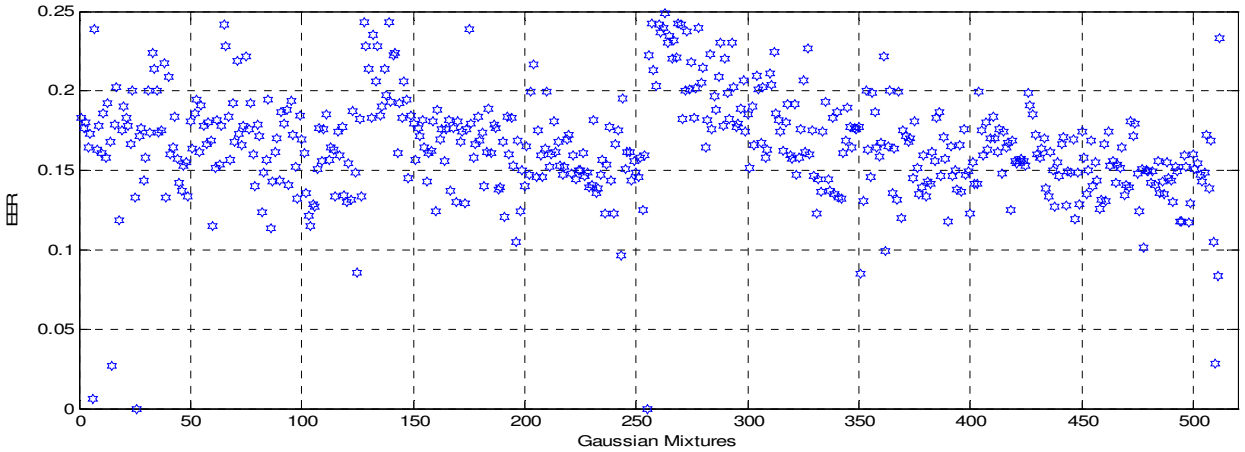

Figure 1. Speaker recognition performance for frame-based LLR corresponding to each Gaussian mixture.

II. SCORE WEIGHTING

A. *Background*

Early work found that there is an exact correspondence between the best scoring frames of the GMMs and particular sounds occurring in the speech [3]. The discrimination between speakers provided by the GMM was in fact due to a subset of phonetic classes. Even though the sounds were not explicitly modeled by a GMM, the scoring appeared to be closely related to them. Since finding an accurate phone recognizer is a challenge in itself, these experiments were repeated in this paper using the best-matched Gaussian mixture for each frame, rather than finding the positions of phonemes. The frame likelihood scores generated by a particular best- matched Gaussian mixture were pooled separately and the EER was obtained by target and impostor scores[1] for that particular Gaussian mixture component. Figure 1 shows the EER of the frames compared to their best matched Gaussian mixtures on male speakers of NIST 2002 Dataset. This shows that those scores generated by certain Gaussian mixtures perform significantly better than others, and would correspond to those phonemes providing the most discrimination among speakers in other systems [3].

B. *Relative Information Level*

The quantity of adaptation data for a particular mixture component determines the accuracy of the MAP adaptation in that region of feature space. The number of occurrences for a particular mixture component in training data, also known as the mixture component soft count [11], has been used as a measurement of the accuracy level of modeling a specific region of the feature space. However, the assumption that the background model contains the same soft count for all mixture components is not always true.

Therefore, a more specific measurement called relative information level, which compares the soft counts of the individual Gaussian mixtures in training and background dataset, is introduced here. The relative information level is derived by decomposing the log-likelihood ratios for each Gaussian mixture. If n_{tarj}, n_{ubmj} are counts of occurrence of the j^{th} Gaussian mixture gm_j in the target and background models, of total length N_{tar}, N_{ubm} frames, respectively, then maximum likelihood estimates of Gaussian mixture probabilities are given by

$$\hat{P}(gm_j \mid tar) = \frac{n_{tarj}}{N_{tar}} \tag{1a}$$

$$\hat{P}(gm_j \mid ubm) = \frac{n_{ubmj}}{N_{ubm}} \tag{1b}$$

Then, the additional information of the j^{th} Gaussian mixture gm_j for a target speaker can be measured by the logarithm of the maximum likelihood estimates

$$u_j = \log \frac{\hat{P}(gm_j \mid tar)}{\hat{P}(gm_j \mid ubm)} \tag{2}$$

The maximum additional information due to the j^{th} Gaussian mixture component of target model versus the same component of background model over all observations can be defined as:

$$I_j = \hat{P}(gm_j \mid tar) \log \frac{\hat{P}(gm_j \mid tar)}{\hat{P}(gm_j \mid ubm)} \tag{3}$$

This *relative information level* I_j provides a means of ranking Gaussian mixtures in terms of speaker modeling accuracy and effectiveness in different regions of feature space. The larger I_j, the more accurate is the j^{th} region of the feature space.

[1] The impostor scores were obtained from 62 target male speakers of NIST 2001 Dataset which were not present in NIST 2002.

C. *Frame-Based Score and Mixture Information Ranking*

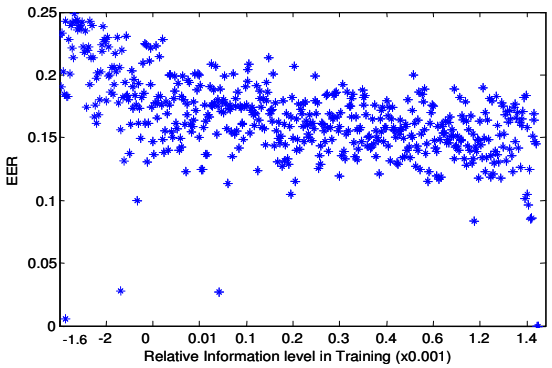

Figure2. Speaker recognition performance versus the information ranking of Gaussian mixtures on target male speakers of NIST 2002 dataset.

Figure 2 shows the error rate for those scores best matched to a particular Gaussian mixture versus the relative information level of that particular Gaussian mixture obtained in training. While there is a strong correlation between the performance of the system and the information level of a Gaussian mixture component (i.e. the Gaussian mixtures of a model that are updated accurately result in better performance) there is also a marked difference in equal error rates between the frames modeled by Gaussian mixtures with the same level of information. This means that the lack of training data is not the only explanation for poor EER. Some frames are more discriminative due to the discrimination ability of their corresponding Gaussian mixtures or more generally their corresponding phonemes.

III. EVALUATION

A. *Database*

Speaker recognition experiments were conducted on cellular telephone conversational speech from the switchboard corpus, the set defined by NIST for the 1-speaker cellular detection task in the 2002 Speaker Recognition Evaluations (SRE). The 2002 set contains 330 targets (139 males and 191 females) and 3570 trials (1442 males and 2128 females) with a majority of CDMA codec utterances; these were scored against roughly 10 gender-matched impostors and the true speaker. The 60 development speakers (2 minutes of speech for each of 38 males and 22 females) and 174 target speakers (2 minutes of speech for each of 74 males and 100 females) from NIST-2001 were used to train the background model of NIST-2002 system. 2038 evaluation test segments (850 males and 1188 females) of NIST-2001 were used as a development dataset to find the best Gaussian mixture sub-set size for male and female speakers.

B. *Baseline System*

The feature set consisted of 15 Mel-PLP cepstrum coefficients [10], 15 delta coefficients plus the delta-energy estimated on the 0-3.8 kHz bandwidth. Cepstral mean subtraction and variance normalization were applied to each speech file during training and testing. The speech detector discarded the 15-20% lowest energy frames before the extraction process. Speaker modeling was based on a GMM-UBM approach [9]. The UBM consisted of two gender-dependent models with 512 Gaussians, trained on 112 male and 122 female speakers from the training portion of development and evaluation datasets of NIST 2001 [10], about 6 hours of data in total. For each target speaker, a GMM with diagonal covariance matrices was trained using the speaker training data via maximum a posteriori (MAP) [3] adaptation of the Gaussian means, with 3 iterations of the EM algorithm.

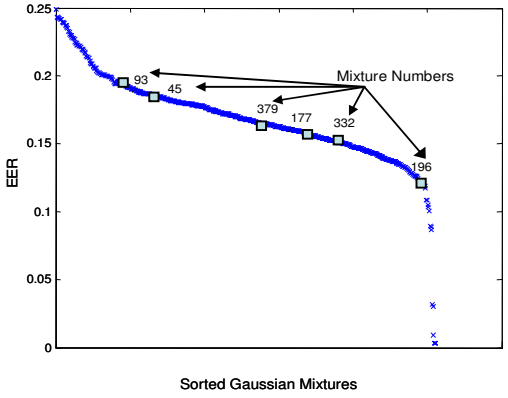

Figure3. Speaker recognition performance for sorted frame-based LLR corresponding to each Gaussian mixture on target male speakers of NIST 2002 dataset

IV. EXPERIMENTAL RESULTS

A. *Weighting Gaussian mixture Scores*

Figure 3 shows that using the frames generated by certain Gaussian mixture components leads to degradation in the performance, e.g. the likelihood frame scores generated by the 93rd and 45th mixture components show a performance of more than 17% EER, whereas using likelihood frame scores generated by more discriminative Gaussian mixtures results in better performance, e.g. the 196th mixture component has about 13% EER. Therefore it would be possible to achieve a better performance by considering those Gaussian mixtures which provide more discrimination between speakers. This confirms previous findings that using a subset of most discriminative phonemes results in a better performance than using all the phoneme classes [3].

B. *Selecting an Optimum Subset of Gaussian Mixtures*

Figure 4 shows results of experiments based on the same methodology as [3], but for Gaussian mixtures instead of phonemes. Here the scores of the frames with *n*-best Gaussian mixtures are summed with equal weight to produce the final score, which is evaluated in terms of EER and minimum DCF. The dashed line shows the variation of minimum DCF by increasing the sub-set size of discriminative Gaussian mixtures. The dotted and dotted-dotted-dashed horizontal lines show the baseline performance achieved by the system using all the frames from different region of feature space, scored in terms of EER and minimum DCF, respectively. It is evident (Figure 4) that a limited number of mixture components are responsible for the whole performance of speaker verification system. It can also be seen that a carefully selected subset of 400 to 420 mixtures provides better results than the standard system using all frames from all Gaussian mixtures, in terms of both EER and min DCF. It is remarkable (Figure 4) that fewer Gaussian

mixtures are responsible for the performance of minimum DCF compared with EER as it almost reaches the optimum minimum DCF with smaller subset size. These empirical results for both metrics indicate that for 512 mixtures, the 100 (approximately) least discriminative Gaussian mixtures (the 420th to 512th least discriminative) degrade the whole

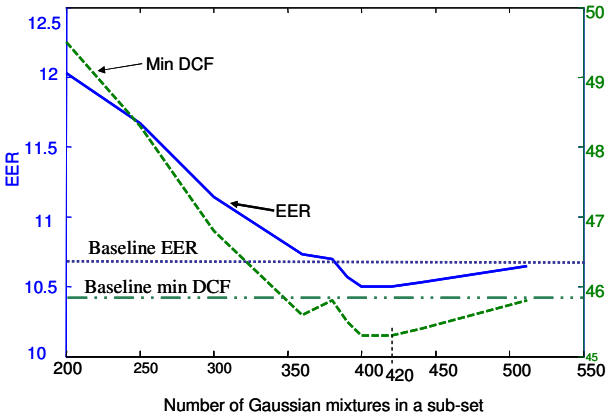

Figure 4. Speaker recognition performance for male speakers using the subset of frames best matched to the most discriminative Gaussian mixtures of the development dataset.

performance and should be omitted from final scoring.

The GMM system combines all frame-based likelihood scores equally, regardless of their ability to discriminate speakers. As has been shown in figure 4, it is possible to consider only a subset of frame-based likelihood scores which correspond to the most discriminative Gaussian mixtures. Although a more complicated weighting method for obtaining the final score would result in a better performance, a simple equal weighting has been used just to show the importance of de-emphasizing or ignoring the frames with low discriminative ability.

Figures 5 shows the Detection Error Tradeoff (DET) curve for the baseline and equal weight subset frame selection

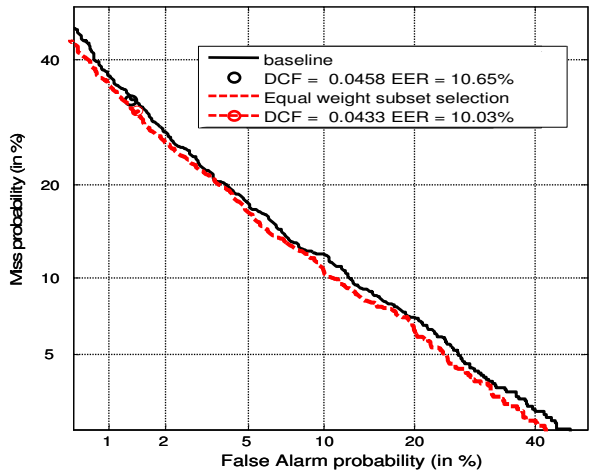

Figure 5. DET curve for the baseline and weighted score systems for male speakers of the NIST 2002 dataset.

technique on male and female speakers respectively.

It can be clearly seen in figure 5 that the equal weight subset selection method performs better than the baseline at the EER operating point, in the area of minimum DCF and in low miss-rate areas on male speakers.

Generally, this technique was more successful for male speakers than for the female population, which concurs with previous results obtained for subset phoneme selection on female speakers [3].

V. CONCLUSION

This paper has presented a comparison between the contributions of different regions of feature space to the final performance of the system. It has been shown that choosing the subset of best-matched Gaussian components corresponding to the frames which resulted in lower error rates during training can improve the error rate of the system. An equal weight has been used to find the final likelihood score. However, it is possible to use more complicated nonlinear methods to estimate the weights to improve the performance, and also comparing this technique to the discriminative training will be investigated in future work. The more recent NIST Databases will be used for future work.

ACKNOWLEDGEMENT

Funding for this research was fully provided by National ICT Australia (NICTA).

REFERENCES

[1] M. J. Carey, E. S. Parris and J. S. Bridle, "A speaker verification system using Alpha-Nets," in *Proc. IEEE ICASSP*, 1991, pp.397-400.

[2] K.-P Li, and J. E. Porter, "Normalization and selection of speech segments for speaker recognition scoring," in *Proc. IEEE ICASSP*, vol. 1, 1988, pp. 595-598.

[3] R. Auckenthaler, E. Parris, and M. Carey, "Improving a GMM speaker verification system by phonetic weighting," in *Proc. IEEE ICASSP*, 1999, pp.313-315.

[4] M. J. Carey, E. S. Parris, S. J. Bennett and H. Lloyd-Thomas, "A comparison of model estimation techniques for speaker verification," in *Proc. IEEE ICASSP*, 1997, pp. 1083-1086.

[5] J. L. Gauvain, and C. Barras, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chain," *IEEE Trans. Acoust,. Speech Signal Process.,* vol.2, 1994, pp.291-298.

[6] M. Nosratighods, E. Ambikairajah, J. Epps, M. Carey, "A novel technique for the selection of speech segments for speaker verification," in *Proc. Int. Conf. Speech Science Tech.*, 2006, pp. 136-141.

[7] M. Nosratighods, E. Ambikairajah, J. Epps, M. Carey, "P-value segment selection technique for speaker verification," in *Proc. IEEE ICASSP,* 2007, pp.269-272.

[8] J. H. Wright, M. Carey, E. S. Parris, "Statistical models for topic identification using phoneme substring", in *Proc. IEEE ICASSP,* 1996, pp.307-310.

[9] M. J. Carey, E. S. Parris, S. J. Bennett and H. Lloyd-Thomas, "A comparison of model Estimation techniques for speaker verification," in *Proc. IEEE ICASSP*, 1997, pp.1083-1086.

[10] J. L. Gavain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, 2002, pp. 89-108.

[11] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing,* vol.10, no. 1/2/3, pp.19-41, 2000.