Effectiveness of speaker-dependent feature score pruning in speaker verification

Surosh G Pillay and Aladdin Ariyaeeinia
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK
{s.g.pillay, a.m.ariyaeeinia}@herts.ac.uk

Mark Pawlewski
BT Labs
Martlesham, Ipswich, IP5 7RE, UK
mark.pawlewski@bt.com

***Abstract*— This paper presents investigations into the use of speech feature score pruning for enhancing the speaker verification accuracy. A new technique based on defining speaker-specific regions for discarding feature scores is proposed and experimentally investigated. The scope of the investigations covers both text-dependent and text-independent speaker verification. Based on the results, it is shown that considerable improvements can be obtained in the former scenario. The paper discusses the motivation for the proposed approach and details the experimental study.**

Keywords- speaker verification, feature score pruning, score normalisation

I. INTRODUCTION

The aim of an automatic Speaker Verification (SV) system is to decide whether to accept or reject a claimed identity based on a given test utterance. Such systems can be text-dependent, where the user is constrained to a fixed utterance (e.g. password) or text-independent, where the verification process is independent of the spoken utterance. In recent years, a popular technique for tackling this problem has been that based on the use of speaker dependent Gaussian Mixture Models (GMM). These speaker models are adapted from a Universal Background Model (UBM) using the Maximum A-Posteriori (MAP) method [1]. In the GMM-UBM approach, the measure of similarity between the given test utterance and the reference model is expressed in terms of log-likelihood ratio (LLR). This is computed by first averaging the log-likelihoods obtained for the individual test feature vectors against the target model, i.e.

$$\log p(X \mid \lambda_T) = \frac{1}{T} \sum_t \log p(x_t \mid \lambda_T) \qquad (1)$$

where X is the test utterance represented by a sequence of feature vectors $\{x_1, x_2, x_3, \ldots, x_T\}$, p represents the probability (likelihood), and λ_T is the target model. The resultant log likelihood for the full test utterance is then normalised by the log likelihood obtained for the test utterance against a UBM.

$$LLR = \log p(X|\lambda_T) - \log p(X|\lambda_{UBM}) \qquad (2)$$

where λ_{UBM} is the universal background model (UBM). This normalised score is subsequently used as the basis for making a binary decision for accepting or rejecting the claimant.

Such score normalisation has been shown to help alleviate the effects of noise contamination across the whole test utterance [2, 6, 7]. However, under practical conditions, the degradation is normally non-uniform. As a result, some feature vectors are more severely contaminated than others. In order to increase the effectiveness of speaker verification in these scenarios, it is logical to consider discarding the least reliable feature scores. To achieve this, various approaches based on pruning of the outlier feature vector scores from the computation of the overall LLR score [3-5] have been proposed. This paper presents further investigations in this area, based on an alternative pruning approach. The proposed technique involves defining speaker-specific regions for discarding feature vector scores.

The remainder of the paper is organised as follows. Section II discusses the motivation behind the use of the proposed approach. The experimental investigations and analysis are given in Section III, and the overall conclusions are presented in Section IV.

II. PROPOSED APPROACH

An essential requirement of any feature pruning technique is the ability to accurately and effectively select the feature vectors which yield uncharacteristically low likelihood scores. This problem can be due to a number of factors such as adverse operating conditions or speaker generated variations. Regardless of the cause, the net result is the reduction of verification score. This in turn leads to high verification errors and incorrect decisions in discriminating amongst individuals in practice. To date, a number of techniques such as Missing Feature Theory (MFT) [3] and frame pruning [4] have been proposed to address this issue. The former concentrates on discarding feature vectors affected by noisy operating conditions while the latter focuses on emphasising high scoring vector scores in the LLR computation. In [5] another approach, using a selection mechanism for feature vectors based on multiple speaker models, is proposed. The method reported in this paper is based on identifying the speaker-dependent regions for pruning undesired feature scores. The limits of such speaker-specific regions are defined by the relative usefulness of the feature scores. To obtain a meaningful comparison of the quality of feature vector scores, a normalisation technique must be applied beforehand.

978-1-4244-1688-2/08/$25.00

As indicated earlier, one such technique that has shown to be effective is UBM which is incorporated in GMM-UBM [6]. Fig 1 illustrates the concept of deploying the proposed method for determining the feature score pruning region for each registered speaker.

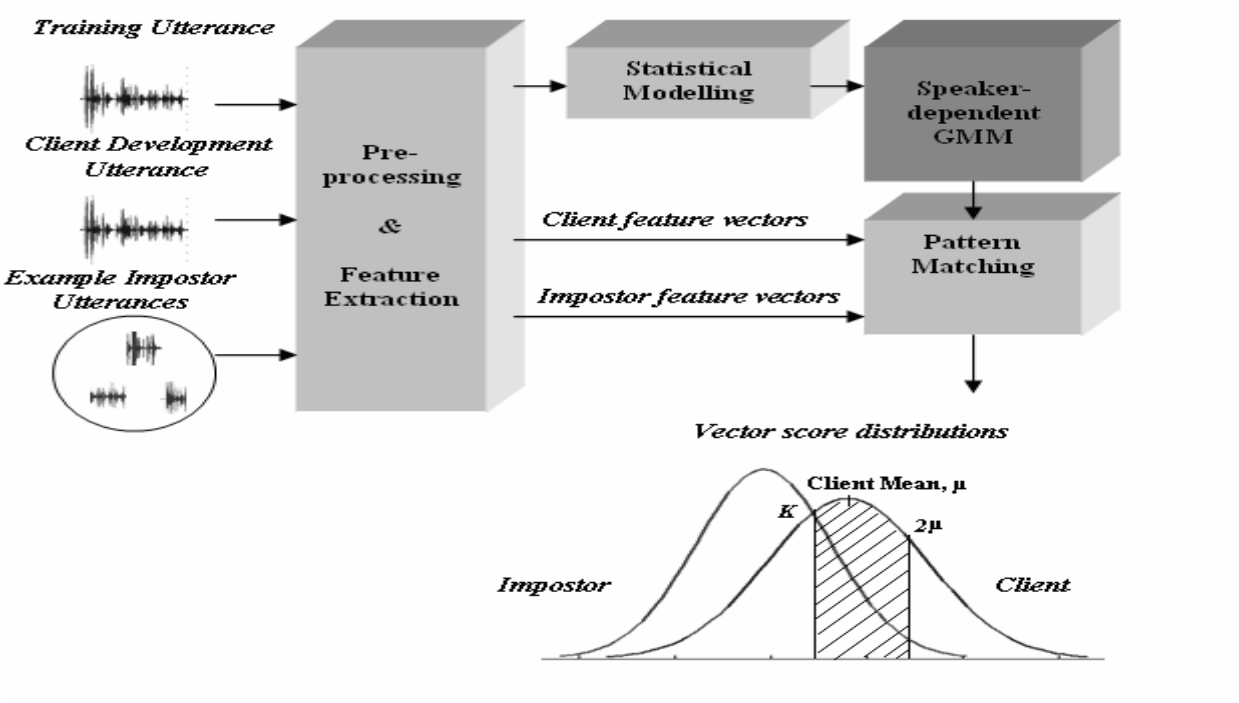

Figure 1. Approach to determining speaker-dependent pruning regions

As illustrated in this figure, first a speaker dependent GMM is obtained from the training utterance using the conventional GMM-UBM with MAP adaptation. A client development utterance is then tested against his/her model, and the resultant vector scores are stored. Example impostor utterances are then tested against the target model to obtain a set of impostor vector scores. The client and impostor distributions of vector scores are then used to empirically determine a speaker-dependent feature discarding region. This region is defined using such parameters as the impostor score mean, client score mean and the intersection point k, given as

$$\kappa = \frac{\sigma_I \mu_C + \sigma_C \mu_I}{\sigma_I + \sigma_C} \qquad (3)$$

where μ_C and σ_C are mean and standard deviation for the client feature score distribution, and μ_I and σ_I are the corresponding parameters for the impostor feature score distribution.

In the testing phase, The UBM-based normalisation is applied at the feature level. The pruning region is then determined based on the speaker specific parameters obtained for the target model in the development stage. This allows a decision to be made about the feature scores to be discarded. The LLR score is then obtained as the average of the log of the remaining feature vector scores. In order to further enhance the performance of the proposed approach, an additional score normalisation such as Unconstrained Cohort Normalisation (UCN) can be applied to the final score [8]. Fig. 2 illustrates the procedure for discarding vectors in the test stage.

It should be pointed out that the main concern in this study is text dependent speaker verification which is currently more viable for commercial applications than text-independent speaker verification. However, for completeness, the effectiveness of the proposed method is investigated for both text-dependent and text-independent verification.

III. EXPERIMENTAL INVESTIGATIONS

A. Speech Data

The speech database used for purpose of text-dependent experiments is the degraded XM2VTS [9]. 99 speakers are enrolled on the system using one training utterance of about 4 seconds for each one. One development and two test utterances of similar duration from different sessions are then used for the respective stages. The example impostor utterances for the development stage are obtained from within the set of registered speakers. For the purpose of testing, 95 out-of-set impostors are included resulting in 198 client scores and 28,809 impostor scores. A UBM for model adaptation and normalisation is trained using a subset of 100 utterances from speakers other than the ones registered or used as the out-of set impostors.

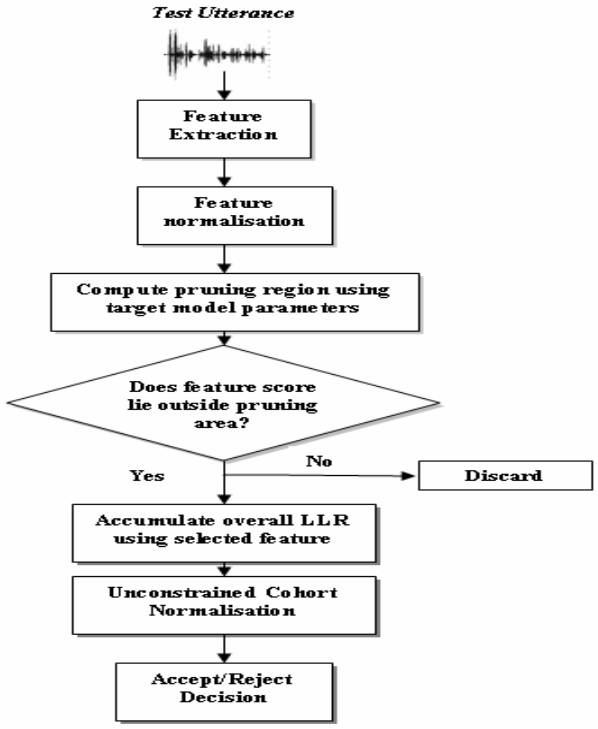

Figure 2. Operations involved in the proposed approach

The dataset used in the text-independent case, is a subset of the 1-speaker detection task of NIST Speaker Recognition Evaluation 2003. For the registered client set, 99 speakers are randomly chosen and modelled using about 2 minutes of speech for each one. As in the text-dependent setup, for each client one development utterance and two test utterances of up to 60 seconds are used. For testing purposes, 95 out-of-set impostor utterances are used, resulting in the same number of client scores and impostor scores as in the previous setup. The UBM in this case is trained using about 8 hours of speech from a subset of 100 speakers [7]. The speakers used for UBM are other than those used as clients or out-of set impostors.

B. Feature Representation

For the purpose of this study, the t^{th} frame of the input speech data is represented as $\mathbf{c}_t \equiv \{c_t(1), c_t(2),\ldots, c_t(20), \Delta c_t(1), \Delta c_t(2),\ldots, \Delta c_t(20)\}$, where $c(i)$ is the i^{th}, mean subtracted, linear predictive coding-derived cepstral (LPCC) parameter and $\Delta c(i)$ is the i^{th} delta LPCC parameter. The extraction of LPCC parameters is based on first pre-emphasising the input speech data using a first order digital filter and then segmenting it into 30 ms frames at intervals of 15 ms using a Hamming window. $\Delta c_n(i)$ was generated by fitting a linear regression line to $c_{t-2}(i)$, $c_{t-1}(i),\ldots, c_{t+2}(i)$.

C. Speaker Representation

In all the experimental investigations discussed in this paper, the speaker representation is based on Gaussian mixture models (GMM). Each speaker model is adapted from a 128m or 2048m, gender-independent UBM using MAP adaptation, in the text-dependent and text-independent setups respectively. The Gaussian mixture densities, m, are parameterised with mean vectors and diagonal covariance matrices.

D. Development and Testing Procedure

The procedure for obtaining the speaker dependent parameters during the development stage is detailed in the previous section. Once these parameters are found for each speaker, pilot experiments are carried out to find the optimum pruning range. Table 1 shows the results obtained for various score pruning ranges in terms of equal error rate (EER). It should be noted that the text-dependent setup is used for the purpose of this range evaluation.

Interestingly, it is seen from the results in Table 1 that extending the pruning criterion to the very low client scores degrades the accuracy. On the other hand, it is found that discarding high impostor scores results in better performance.

TABLE I. RESULTS OBTAINED USING VARIOUS PRUNING RANGES

Pruning Range (x) Description	EER (%)
Baseline GMM-UBM	3.67
2× Impostor Mean < x < k	3.94
Impostor Mean<x < k	3.92
Impostor Mean+ k < x < Client Mean-k	3.03
Impostor Mean < x < Client Mean	3.03
2k<x<Client Mean	3.03
k<x<Client Mean	2.81

According to these results the optimum range for feature score pruning is between "the intersection point of the impostor and client vector score distributions" and "the client score mean". However, based on further investigations it is found that the best performance is obtained when the pruning is extended to the right of the client score mean (Fig. 3), i.e.

$$k < x < \alpha . \mu_C \quad (4)$$

where μ_C has the same meaning as in (2) and α needs to be chosen empirically from the development set. As seen in Fig. 3, for the considered database, the optimum range of α is between about 1.5 and 2. As expected, large values of α result in high error rates as most of the clients' high vector scores are removed. It is also important to point out that the value of α should not be over-tuned to the development data as this may affect the results in the test phase (only intervals of 0.5 should be considered).

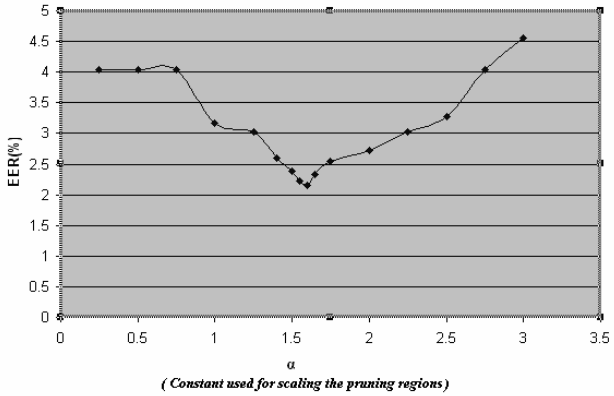

Figure 3. Error rate as a function of α

E. Text-dependent Results and Discussions

The aim of the first set of experiments is to investigate the effectiveness of using the pre-defined pruning region, in a text-dependent scenario. Table 2 shows a comparison of the results obtained using the proposed technique and the standard GMM-UBM. The evaluation also includes results obtained with GMM-UBM plus UCN (since UCN is used to obtain the final LLR score in the new approach). Figure 4 further illustrates the results obtained in this part of the study as DET (detection error rate trade off) plots.

TABLE II. COMPARISON OF THE NEW TECHNIQUE WITH BASELINE METHODS IN A TEXT-DEPENDENT SCENARIO

Method	EER (%)
GMM-UBM	3.54
GMM-UBM & UCN	2.25
Proposed technique	1.63

It is observed from Table 2, that the proposed pruning technique outperforms both baseline methods by relative improvements of 53% and 27.5% respectively. It is also seen from the DET plots that the increased accuracy is obtained over all operating regions except where a high False Alarm probability is compromised for a low Miss Probability.

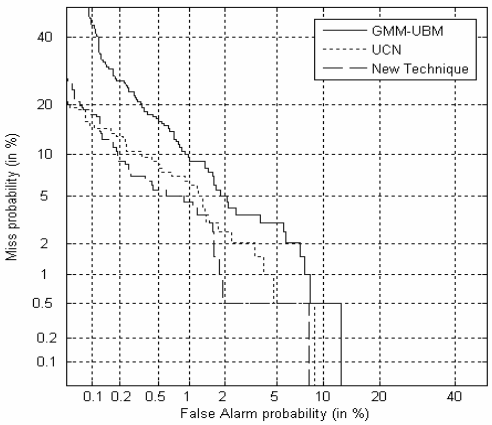

Figure 4. DET plots for the text-dependent experiments

F. Text-independent Results and Discussions

The next set of experiments evaluates the effectiveness of the proposed approach in text-independent speaker verification. As before, the baseline results are based on using GMM-UBM with and without UCN.

In this scenario, it is seen (Table III) that the relative increases in the accuracy obtained with the proposed approach over the baseline methods are 21.5% and 9.8%. According to Table III and Fig. 5, in this setup, the improvements are lower than those obtained in the previous section. This is believed to be due to the effects of variation in unseen data between the development and testing stages. Nevertheless, the DET plots and the EER values indicate that the proposed method can still be of value in text-independent verification.

TABLE III. COMPARISON OF THE ACCURACY OF THE PROPOSED TECHNIQUE WITH THOSE OF BASELINE METHODS IN TEXT-INDEPENDENT EXPERIMENTS

Method	EER (%)
GMM-UBM	11.50
GMM-UBM & UCN	10.00
Proposed technique	9.02

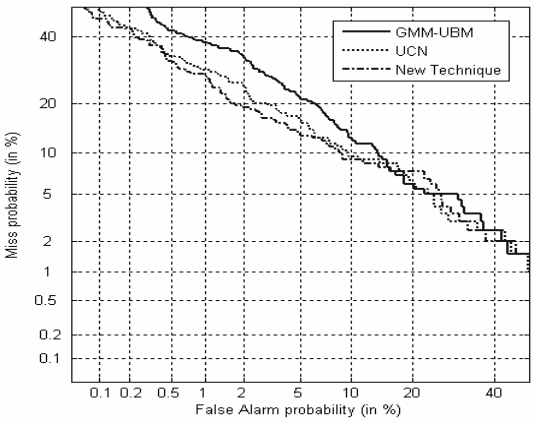

Figure 5. DET plots for the text-independent experiments

IV. CONCLUSION

Pruning feature vector scores in GMM-UBM approaches is an area which has shown promising prospects in speaker recognition. In this paper, a new method for feature score pruning in speaker verification is presented and investigated. The proposed approach involves defining speaker-specific regions for discarding feature vector scores. Based on the experimental results, it is shown that the proposed score pruning method can considerably improve the accuracy of text-dependent speaker verification. In the text-independent case, whilst some accuracy improvement has been achieved, this has not been as extensive as that for text-dependent. This is thought to be due to the effects unseen data in text-independent process. In particular, the effects of variations in unseen data between the development and testing stages will need to be further investigated. The projected work in this area also includes investigating the effectiveness of the proposed score pruning method with SVM (support vector machine)-GMM-based speaker verification [10].

V. REFERENCES

[1] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian Mixture Models”, Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems”, in Digital Signal Proc, vol. 10, pp. 42-54, 2000.

[3] A. Drygajlo and M. El-Maliki, “Speaker verification in noisy environments with combined spectral subtraction and missing feature theory,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98), vol. 1, pp. 121–124, Seattle, Wash, USA, May 1998

[4] L. Besacier and J.F. Bonastre, "Frame pruning for speaker recognition". In Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Seattle, USA, May 1998

[5] S. Kwoon and S.Narayanan, “Robust speaker identification based on selective use of feature vectors”, Pattern Recognition Letters, vol. 28, Issue 1, pp. 85-89, 2007.

[6] A. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, “Verification effectiveness in open-set speaker identification, in IEE Proceedings Vision, Image and Signal Processing, vol. 153, No. 5, October 2006, pp 618-624

[7] D. Reynolds, “Comparison of background normalisation methods for text-independent speaker verification”, in Proc. Eurospeech 1997, Rhodes, pp. 963-966, 1997.

[8] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, A. Malegaonkar, “Relative effectiveness of score normalization methods in open-set speaker identification”, in Proc. Speaker Odyssey, pp. 369-376, 2004.

[9] L.Besacier, P.Mayorga, J.F.Bonastre, C. Fredouille and S. Meigner, “Overview of compression and packet loss effects in speech biometrics”, Biometrics on the internet, IEE Proc.-Vis. Image Signal Process., Vol. 150, No. 6, December 2003

[10] W.M. Campbell, D.E. Sturim and D.A. Reynolds, “Support Vector Machines using GMM supervectors for speaker verification”, Signal Processing Letters, IEEE, vol. 13, Issue 5, 2006.