TWO SPEAKER DETECTION BY DUAL GAUSSIAN MIXTURE MODELLING

S. Myers, J. Pelecanos and S. Sridharan
Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, George St, Brisbane, AUSTRALIA, 4001.
sd.myers@qut.edu.au, j.pelecanos@qut.edu.au
s.sridharan@qut.edu.au

ABSTRACT: Presented in this paper is a method for performing two speaker detection utilising a technique of modeling the output scores of an Adapted Gaussian Mixture Model – Universal Background Model (GMM-UBM) system. This method consists of training a two mixture Gaussian Mixture Model on the output scores of a speaker recognition engine. A baseline system developed for the NIST 2000 Speaker Recognition Evaluation has demonstrated encouraging results, which are presented. The computation time of this system was significantly less when compared to some of the systems submitted for the NIST 2000 competition. Improvements to the baseline system are suggested, and current experiments indicate that two speaker detection systems based on this method will demonstrate very good performance.

1.0 INTRODUCTION

Two speaker detection is the task of determining whether or not a specified target speaker is talking in a conversation consisting of two people. The two speaker detection system proposed in this paper is presented in the context of the National Institute of Standards and Technology (NIST) 2000 Speaker Recognition Evaluation (NIST, 2000) in which the authors participated. The NIST evaluation is an annual international event aimed at improving the state-of-the-art technology in speaker recognition. Since 1998, the evaluation has included a two speaker detection task. This is an extension of the one speaker task, which is the basic recognition task for all evaluations since 1996.

Two speaker detection is more difficult than one speaker detection for a number of reasons. Firstly, many of the channel estimate/reduction techniques used to improve the performance of one speaker detection are rendered useless because the speech signal present is sourced from two different handsets and/or channels. Secondly, in any conversation, there are inevitably times when both participants are speaking at the same time, a phenomenon known as double talk. Double talk can have an understandably detrimental effect on multi-speaker recognition technologies.

For the purposes of the NIST Speaker Recognition Evaluation, it was desired that a two speaker detection system be developed that could be adapted quickly and easily from the one speaker detection engine. This was done so that most of the development time could be centrally focussed, requiring only a small commitment of time to adapt a high performing two speaker detection system. The technique described in this paper reflects this objective.

The following section describes the core engine that was used for both the one and two speaker detection systems, while the two speaker subsystem, which forms the focus of this paper, is discussed in section 3. Results obtained from the NIST 2000 Evaluation are presented in section 4, with the conclusions and discussion presented in section 5.

2.0 GMM SPEAKER ADAPTATION SYSTEM

The two speaker detection system discussed in this paper is largely based on the one speaker detection engine that was used in the NIST 2000 Speaker Recognition competition. This engine is derived from the Adapted GMM-UBM scheme (Reynolds, 1997), which has been used successfully in several NIST competitions. The model adaptation process requires the training of a high order GMM on a large quantity of speech. A GMM is a linear combination of $k = 1, 2, ..., N$, single Gaussian components of dimensionality D, with mixture weights p_k, means $\vec{\mu}_k$, and diagonal covariance matrices Σ_k. For a single speech feature vector observation, $\vec{X}$, the probability density function for a speaker model λ, is described.

$$p(\vec{X} \mid \lambda) = \sum_{k=1}^{N} p_k \; g(\vec{X}, \vec{\mu}_k, \Sigma_k) \tag{1}$$

with

$$g(\vec{X}, \vec{\mu}_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} \exp\left[-\frac{1}{2} (\vec{X} - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{X} - \vec{\mu}_k) \right] \tag{2}$$

Two UBM's, one male and one female were generated for the evaluation. These models were generated using electret handset data from the NIST '99 evaluation, with each UBM comprising 512 Gaussian mixtures. The UBM's were initially estimated using 10 iterations of the VQ seeding algorithm (Pelecanos, 2000), with the final models being generated after 40 iterations of the Expectation Maximisation algorithm. Target speaker models were adapted from the appropriate UBM toward the target speech vectors using MAP adaptation (Gauvain, 1994). The feature vectors used for training and testing were 24 dimensional vectors consisting of 12 Mel-Frequency Cepstral Coefficients (MFCCs) with their corresponding delta coefficients. The MFCCs were derived from a bandlimited (300-3200Hz) signal; using 32ms frames produced at 10ms time intervals. An energy-based scheme was used to remove silence frames.

To compensate for channel effects, a novel sliding cepstral mean removal technique was employed. This technique uses a 300-frame window to produce an estimate of the channel effects, the value of which is subtracted from the frame in the middle of the window. The window is then slid along by one frame and the process is repeated. This technique is applied to each of the MFCCs before the delta coefficients are generated. The sliding-mean removal technique gives similar results to cepstral mean subtraction alone for speaker detection, and is far simpler to implement and faster to compute than some RASTA (Van Vuuren, 1999) based technologies. Segment based cepstral mean subtraction should not be used for the two speaker detection task because of the multiple channel sources. This makes the sliding window technique particularly valuable.

The speaker models generated for the one speaker task were used with no modification in the two speaker task. These models were generated using approximately two minutes of speech from a single electret handset telephone conversation. The scoring of the speaker models during testing is where the one and two speaker detection tasks differ, and this is discussed.

3.0 SPEAKER SCORING

Speaker scoring is the final stage in this speaker recognition system, and involves classifying a given test segment against a specified target model. If the score after classifying is above a pre-determined threshold γ, then the test segment is identified to have come from that target model. More formally, this is represented by a hypothesis test in the form of a likelihood ratio score of the test speech segment X, given a target model λ_{TARGET}, and a UBM λ_{UBM}.

$$\frac{p(X \mid \lambda_{TARGET})}{p(X \mid \lambda_{UBM})} \underset{\text{Impostor}}{\overset{\text{Target}}{\gtrless}} \gamma \tag{3}$$

This comparison may also be specified in log form.

$$\log p(X \mid \lambda_{TARGET}) - \log p(X \mid \lambda_{UBM}) \underset{\text{Impostor}}{\overset{\text{Target}}{\gtrless}} \log \gamma \tag{4}$$

For both the one and two speaker tasks, the first phase of the scoring process is identical, and involves generating a likelihood score for each feature vector extracted from the test segment. Each vector is compared against the UBM, and the 5 highest scoring mixtures are used to generate a log-likelihood score. The feature vector is then classified against the target model, using the *same* 5

mixture indexes that were used for the UBM (Reynolds, 1997). The UBM log-likelihood score is then subtracted from the target log-likelihood score to generate the final log-likelihood-ratio (LLR) score for the frame. This process is repeated for every non-silence/speech frame in the test segment given by $X = \{\vec{X}_1, \vec{X}_2, ..., \vec{X}_T\}$, producing a final set of T scores denoted by Λ_t.

$$\Lambda_t = \log p(\vec{X}_t \mid \lambda_{TARGET}) - \log p(\vec{X}_t \mid \lambda_{UBM}) \tag{5}$$

For the one speaker detection task, the final step in the process is to obtain the average of the T scores and apply any specific normalisation techniques to this score. An expected LLR score is determined for the feature vectors of a test speech segment in the following modified hypothesis test.

$$E[\Lambda_t] = \frac{1}{T}\sum_{t=1}^{T}\Lambda_t \underset{\text{Impostor}}{\overset{\text{Target}}{\gtrless}} \log \gamma' \tag{6}$$

In contrast, the final phase in the two speaker task is substantially more difficult, because in the final set of T scores $\Lambda_1, \Lambda_2, ..., \Lambda_T$, it is not known which of the scores belong to the first speaker and which belong to the second. There are a number of potential ways to solve this problem.

3.1 External Segmentation

The first possibility for solving this problem is the obvious one: divide the original speech sample into the segments belonging to the first and second speaker (Dunn et al, 2000). Once achieved, the problem is simply reduced back to two separate tests of the one speaker detection task. After processing both speakers' segments separately, the highest score of the two is returned. While this is certainly attractive, there are definite drawbacks. Most importantly in terms of the NIST evaluation is the fact that splitting the original signal into two components accurately can be a slow and computationally expensive operation. However, external segmentation has been shown to have a slight improvement over the internal segmentation approach.

3.2 Internal Segmentation

In the second method, rather than segmenting the two speakers based on the input signal, segmentation is performed using the output log-likelihood-ratio scores only. Internal segmentation works by assuming that the scores from two different speakers will be statistically different from each other, particularly in the case where one of the two speakers is in fact the hypothesised target speaker. By taking the average of the better performing log-likelihood-ratio scores, it is anticipated that this figure of merit will represent the more likely of the two speakers. Score extraction may be achieved by taking those scores above some predetermined threshold, or alternatively taking the N% best scoring frames (NIST Website).

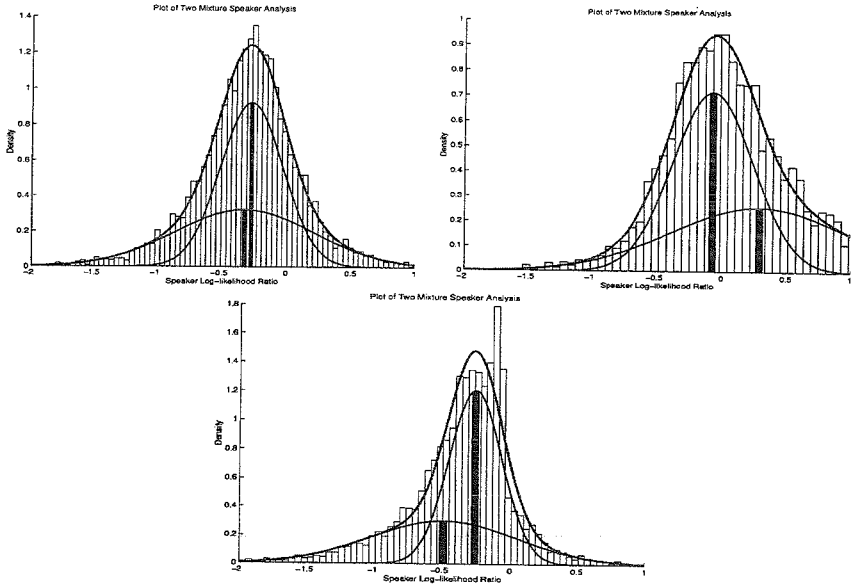
These methods, while certainly quick, are flawed by their simplicity. They are substantially affected by the log-likelihood-ratio score statistics of both speakers present in the test speech. A more consistent approach to internal segmentation is discussed in the next section.

3.3 Bi-Modal Gaussian Distribution Analysis

As the name suggests, this method assumes that in the distribution of the final log-likelihood-ratio scores, each speaker contributes a single Gaussian mode. However, unlike the internal segmentation method described above, no assumption is made about the overlap between the modes. One of the major problems with the previously mentioned internal segmentation method is the fact that a certain percentage of the best scores chosen will almost certainly have come from the unwanted speaker. This effectively taints the final averaged score.

An alternative using a bi-modal analysis method will be described. After the set of N log-likelihood-ratio scores have been evaluated, a two mixture GMM is trained using 20 iterations of the E-M (Expectation Maximisation) algorithm. This is a rapid operation, because the dimensionality of the

GMM parameters is low, and the number of individual scores is also relatively small – approximately 1500-2000 depending on the quantity of silence. (By contrast, training the original UBM involves a 24-dimensional GMM with 512 mixtures, using approximately 300,000 individual feature vectors as training points.) After the analysis is complete, the mode with the highest mean is chosen, and this mean value is used as the score for the final thresholding operation. To illustrate the method, consider the following distributions taken from the NIST 2000 two speaker detection evaluation.



Figures 1 (a), (b) and (c): The figure on the top left (a) is a bi-modal analysis for the case where both speakers are impostors. The figure on the top right (b) is an analysis where one speaker is the hypothesised target speaker. The lower figure (c), is the case where one of the impostor speakers is male, and the other is female. Note that in this case there are more than two modes, which are not modelled appropriately with two mixtures.

The diagrams (Figures 1 (a), (b) and (c)) presented are representative of the impostor/target speaker distribution mix that is possible for the two speaker detection task. In particular, the inability of the system to model the cross gender test case caused a degradation of the final NIST 2000 results, and this is discussed later. The following section presents the results that were obtained for the NIST 2000 Evaluation.

4.0 RESULTS

The standard method of evaluating performance in speaker verification tasks is with the DET curve. A description of the DET curve can be found in (Martin et al, 1997). For the purposes of this paper, it is sufficient to note that a better performing system is one that has lower miss and false alarm probabilities. Figure 2 shows the DET curve of the NIST 2000 evaluation. The proposed system is marked as the *QUT System*. The remaining curves correspond to submissions presented by the other participants of the NIST evaluation. It is worth noting that the proposed system was placed third. The other two systems that were placed ahead of the QUT system used external segmentation, which carried a relatively high computational load. Significant savings in computation times were achieved in the QUT system by using the bi-modal analysis.

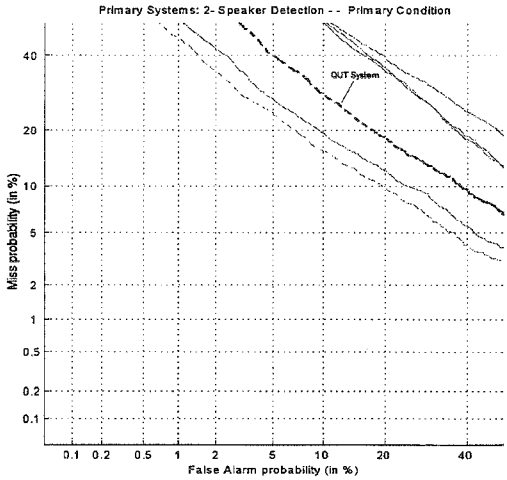


Figure 2: Comparison of all systems that entered the 2-speaker detection task for NIST's primary condition of interest.

Cross gender test segments caused significant problems for the bi-modal distribution in the proposed system. Generally, it is expected that a decision made on a mixed gender test would be of higher accuracy than a same gender test. This is not what is observed for our system. The DET curve in Figure 3 emphasises the potential for improvement of the system. It is hypothesised that the cross-gender tests performed worse because a gender specific UBM was used for speaker modeling.

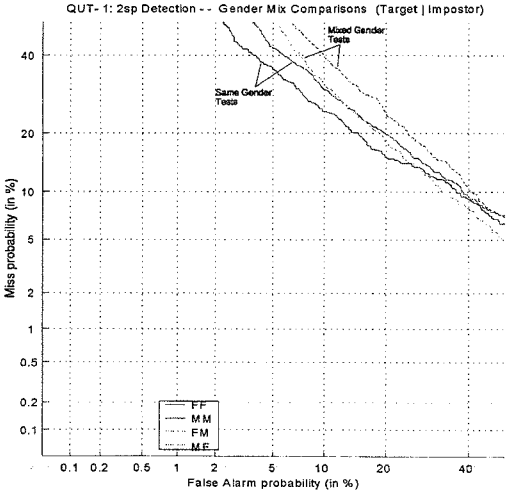


Figure 3: DET curve from the NIST 2000 two speaker task comparing the results for same gender and cross gender trials.

The bi-modal analysis was compared with the internal segmentation method of averaging the top *N*% of log-likelihood-ratio scores. Values of *N* used were 10, 30, 50, 70 and 90. The optimum value found was 30%. The bi-modal analysis method was found to marginally outperform this, particularly in the low-false alarm probability region of interest on the DET curve. It should be noted that trial and error had to be used to find the optimum value, and it is highly likely that in another testing environment, a different optimal value for *N* would be required. This provides an inherent advantage of the bi-modal analysis system.

5.0 CONCLUSIONS

Given the speed and simplicity of the bi-modal distribution analysis, the results from the NIST 2000 evaluation are quite promising. However, to improve the results further, work needs to be performed to improve the cross gender results. The first phase is to change to a gender independent UBM rather than using gender specific UBMs. It is anticipated that this will condition the score distribution significantly and help to ensure that a bi-modal distribution is present for more test instances. It is anticipated that this will provide a performance increase.

One of the methods for improving the one speaker detection results dramatically is through the use of speaker score normalisation such as H-Norm (Reynolds, 1997). The current bi-modal system trialled during the NIST Evaluation did not include any such normalisation. It is anticipated that a method for normalising the log-likelihood-ratio score after bi-modal analysis will also be added to the system.

The bi-modal analysis was found to marginally outperform the top N% internal segmentation system with minimal sacrifice of speed. Additionally, it was a baseline bi-modal system that outperformed the optimised internal segmentation method. After anticipated improvements have been made, it is expected that the bi-modal system will outperform the existing internal segmentation method.

Although current technology using external segmentation provides better performance compared to our proposed bi-modal analysis system, the speed and computation requirements of the external segmentation based systems may be prohibitive for real time applications. The bi-modal distribution system presented in this paper presents scope for real time applications.

6.0 ACKNOWLEDGMENTS

This work was supported by a research contract from the Australian Defence Science and Technology Organisation (DSTO).

7.0 REFERENCES

Dunn R, Reynolds D and Quatieri T. (2000) Approaches to Speaker Detection Tracking in Conversational Speech, Digital Signal Processing, Vol 1/2/3, pp 93-112.

Gauvain J and Lee C. (1994) Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, Vol 2, No 2, pp 291-297.

Martin A, Doddington G, Kamm T, Ordowksi M and Przybocki M. (1997) The DET Curve in Assessment of Detection Task Performance, in proc. of Eurospeech, Vol 4, pp 1895-1898.

NIST. (2000) The 2000 NIST Speaker Recognition Evaluation Plan, Version 1.0.

NIST Website. (2000) NIST's Speaker Recognition Website, http://www.nist.gov/speech.

Pelecanos J, Myers S, Sridharan S and Chandran V. (2000) Vector Quantization based Gaussian Modelling for Speaker Verification, in proc. of ICPR, Paper Number 1219.

Reynolds D, (1997) Comparison of Background Normalization Methods for Text-Independent Speaker Verification, in proc. of Eurospeech, Vol 2, pp 963-966.

Van Vuuren S. (1999) Speaker Verification in a Time-Feature Space, PhD Thesis.