Vector Quantization based Gaussian Modeling for Speaker Verification

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

Abstract

Gaussian Mixture Models (GMMs) have become an established means of modeling feature distributions in speaker recognition systems. It is useful for experimentation and practical implementation purposes to develop and test these models in an efficient manner, particularly when computational resources are limited. A method of combining Vector Quantization (VQ) with single multi-dimensional Gaussians is proposed to rapidly generate a robust model approximation to the Gaussian Mixture Model. A fast method of testing these systems is also proposed and implemented.

Results on the NIST 1996 Speaker Recognition Database suggest comparable and in some cases an improved verification performance to the traditional GMM based analysis scheme. In addition, previous research for the task of speaker identification indicated a similar system performance between the VQ Gaussian based technique and GMMs.

1. Introduction

Gaussian Mixture Models [1] have become an effective means of modeling speaker feature distributions for speaker recognition systems. When evaluating large speaker databases, a fast method of training and testing is required, since the computation time with large quantities of data can become extensive. Proposed is a method termed Vector Quantization Gaussian (VQG) modeling. It is an efficient means of calculating an approximation of the Probability Density Function (PDF) of the speaker features.

Past work in this area investigated the use of VQGMM [2] adaptation. This modeling scheme involves the clustering of speech into regions using Vector Quantization and then training a Gaussian Mixture Model on each sub-space. To test the model, each observation is tested against the Gaussian mixture sub-model corresponding to the most likely sub-space region. In this instance, the testing of vectors used only the information gained from the maximum-likelihood estimate of the GMM sub-model from the closest codevector region. There was no consideration of the contributions of the overlapping densities from adjacent VQ partitions. Thus, speaker information that may be present on the regional boundaries would be lost. However, an advantage of this approach is the marked reduction in testing time attributed to avoiding the necessity to accumulate the likelihood scores from all mixtures.

Another version of this scheme was trialled, but using only a single Gaussian distribution for each partition. In this instance, the maximum Gaussian mixture density was scored in the testing process [3]. This method ignored the contribution of other significant mixture components.

A method is proposed to include the contribution of adjacent regions but using computationally efficient single Gaussian components to establish the model. Here, Vector Quantization is used to separate speech vectors into their corresponding regions and a single multi-dimensional Gaussian is calculated for each. A substitute Gaussian Mixture Model is formed that considers density contribution information from adjacent regions by compiling information available from the mixtures and the number of points in each region. This model, termed the Vector Quantization Gaussian (VQG) is capable of accumulating contributions from adjacent Gaussian mixtures with the advantage of improved training speed. Gaussian Mixture Modeling and the proposed Vector Quantization Gaussian modifications are discussed.

2. Vector Quantization Gaussian Modeling

To understand the structure of a VQG, it is important to detail the standard Gaussian Mixture Model approach. A GMM approximates a probability density function by a combination of N sets of D dimensional Gaussians with mixture weights p_k, means $\vec{\mu}_k$, and covariances Σ_k.

Proceedings of the International Conference on Pattern Recognition (ICPR'00)
1051-4651/00 $10.00 @ 2000 IEEE

The likelihood of a single observation vector, $\vec{X}$, given an *N* mixture GMM, λ, can be determined.

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

where

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

Model estimation is performed using an iterative process called the Expectation-Maximization (E-M) [4] algorithm. Typically, for computation speed, only the diagonal components of the covariance matrices are used. Thus, the orthogonal GMM method considers only the mixture variance information when determining the parameters for finding a maximum-likelihood estimate of the data distribution.

An alternative to the E-M based GMM approach is the VQG. The VQG method is initiated by partitioning the speech feature space into regions (indicated by the set of codevectors) using vector quantization. The speaker space is divided into regions *k=1,2,…,K* by use of the corresponding codevectors. K-Means [5] clustering is used to perform the vector groupings.

Given a set of *S* observation vectors $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_S\}$, a set of K codebook vectors $C = \{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_K\}$ must be selected to minimise the distortion function *D*, specified by:

$$D = \sum_{j=1}^{S} \min_{1 \le k \le K} d(\vec{x}_j, \vec{c}_k) \tag{3}$$

with

$$d(\vec{x}_j, \vec{c}_k) = (\vec{x}_j - \vec{c}_k)^{\mathrm{T}} (\vec{x}_j - \vec{c}_k)$$

The K-Means algorithm produces a natural grouping of the final codevectors. These codevectors can be used for assigning the set of test vectors, *X*, to their relevant codevector defined region, indicated by a codevector region index, $\hat{C}_j$.

$$\hat{C}_j = \arg\min_{1 \le k \le K} d(\vec{x}_j, \vec{c}_k) \qquad j = 1, 2, \ldots, S \tag{4}$$

With the training vectors grouped into their codevector bins, an orthogonal multi-dimensional Gaussian is calculated using the mean and variance statistics from the test vectors in each codevector region. An approximation to the Gaussian Mixture Model is determined by estimating the mixture weights p_k, means $\vec{\mu}_k$, and covariances Σ_k. The mixture weight, p_k, is calculated as the proportion of the total number of test vectors that belong to codevector region *k*. Each GMM vector mean $\vec{\mu}_k$, is assigned its corresponding codevector, $\vec{c}_k$. The diagonal components of the diagonal covariance matrix Σ_k, for each mixture, are calculated from the variances of the vector observations in each codevector region (ie. the non-diagonal components are set to 0).

In the ideal situation, this approximation requires that the feature vectors are relatively sparse and well clustered, with each cluster being close to spherical in distribution. Additionally, VQGs have a similar requirement to that of orthogonal GMMs, in that the features should be uncorrelated.

For many applications, including speaker verification, it is difficult to satisfy these criteria. However, by attempting to match these requirements as much as practical, model estimation errors can be minimised. For example, in the instance when the features vary with each other by an order of magnitude, it is useful prior to the experiment to normalise the data by the standard deviation of each feature dimension.

3. Speaker Recognition Systems

There are two main speaker recognition systems examined in this experiment. These are based on background speaker selection and the use of a Universal Background Speaker Model as proposed by Reynolds [6].

The multiple background speaker selection technique takes a group of maximally-spread-far and maximally-spread-close background speakers and uses them in a likelihood ratio comparison. The general system applied for speaker recognition is indicated in Figure 1.

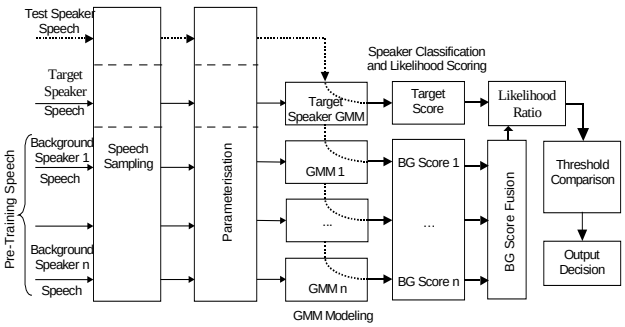

Figure 1. Background Speaker Recognition process.

In contrast, a Universal Background Model (UBM) is a single model trained on masses of speech information from many speakers. The target speaker's model is adapted from information provided by the UBM and the target speaker's speech. For testing, the likelihood ratio using the test speech compared against the target and universal speaker models is calculated.

The Universal Background Speaker Model system can be viewed on similar terms to the Multi-Background Speaker Model scheme. For background speaker scoring purposes, the UBM system can be considered as a multi-background speaker recognition system using a single, large mixture background speaker model.

Given a set of *S* observations $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_S\}$, the mean frame-based log-likelihood can be determined as shown below. This figure is used in the calculation of the model likelihoods in the frame based likelihood ratio calculation.

$$\frac{1}{S}\sum_{j=1}^{S} \log p(\vec{x}_j \mid \lambda) \tag{5}$$

To measure the closeness of a match between a test segment and a target speaker, the likelihood ratio is calculated. It is given as the ratio of the likelihood that a speech segment *X*, is from the target speaker model λ_t, to the likelihood of the same speech segment given that it is not from the target speaker (indicated as $\overline{\lambda}_t$).

$$LR = \frac{p(X \mid \lambda_t)}{p(X \mid \overline{\lambda}_t)} \tag{6}$$

The ratio can be estimated by incorporating the scores from the set of background speaker models. The log-likelihood ratio, given a target speaker model, λ_t, and a set of *B* background speaker models, $\lambda_1, \lambda_2, \ldots, \lambda_B$, is calculated as follows:

$$\log(LR) = \log p(X \mid \lambda_t) - \log\left\{\frac{1}{B}\sum_{b=1}^{B} p(X \mid \lambda_b)\right\} \tag{7}$$

The Universal Background Model log-likelihood ratio calculation is a specific case of this equation with *B* set to one.

4. System Implementation and Evaluation

There were four systems trialled in the experiment. Two experiments were the E-M based GMM scheme with multiple background speakers and with the Universal Background model. The remaining two procedures use the VQG approximation to the E-M based GMM.

The UBM GMM method was established using 512 Gaussian mixtures. This model was trained on the training data obtained from the Development data set of the NIST 1996 recognition data. The background speaker system was formed using 10 background speakers (5 maximally-spread-far and 5-maximally-spread close) with each speaker model comprised of 40 Gaussian mixtures.

All systems used 12 dimensional, 20 filterbank MFCC features with their corresponding delta coefficients. Mean subtraction of the MFC coefficients was performed to reduce channel effects.

These systems were evaluated using data available from the NIST 1996 Evaluation corpus [7]. This database consists of 21 male and 19 female claimant speakers with a selection of approximately 370 impostors. As a general note, the speech data consisted of both matched and mismatched handset recordings with respect to the training handset.

5. Results

The four experiments were evaluated for the complete NIST evaluation and the particular case of one session training and 30 seconds testing for male speakers is presented. Results for other evaluation configurations also revealed a consistent performance trend. There are two plots that indicate the performance of the VQG and standard GMM methods using (i) multiple background speakers and (ii) the Universal Background Speaker Model. Each system performance is shown by three DET curves. In this presentation, the best-performing statistic is when the same training handset is used for testing, and the most degraded result occurs for the non-training handsets. The middle curve is the overall performance that is comprised of the matched and mismatched handset cases.

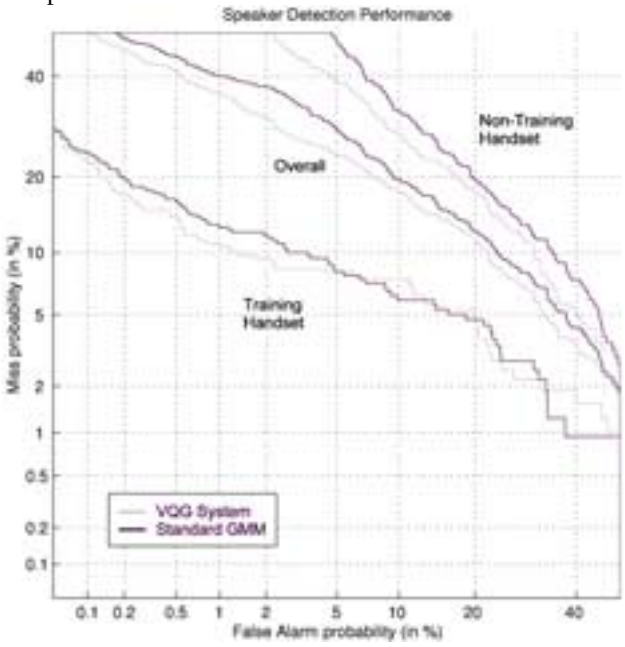

Figure 2. A plot of the VQG and standard GMM DET curve performances using the Multiple Background Speaker technique.

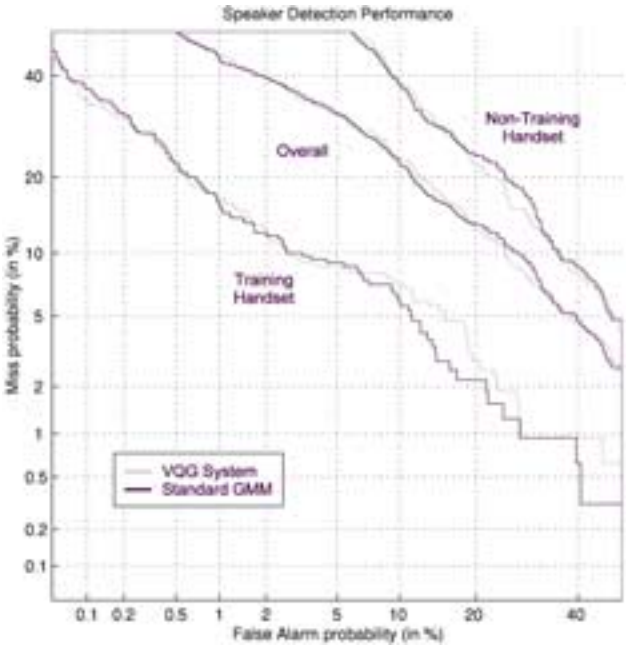

Figure 3. The VQG and standard GMM performances for the Universal Background Model method.

There is a general trend for the VQG scheme to enhance recognition performance for multiple background speakers. This may be attributed to the estimation of the Gaussian means and variances being more robust to mismatched conditions. An improvement in performance is also noted for the matched handset condition. The matched and mismatched handset conditions were determined by NIST by recording the telephone numbers used for each call. A handset classification error will be introduced, as the same phone number will not always imply the same handset.

An analysis of the VQG has indicated that in general, when clusters are distinct, the VQG approximates the standard GMM system well. If there is significant overlap of adjacent mixtures, the VQG scheme tends to form a grouping of mixtures with underestimated mixture variances. With the UBM technique, adjacent mixtures are more likely to overlap and consequently, the mixture variances may be underestimated. Hence, a decrease in performance was observed for the *training* handset over a limited region for the UBM approach. However, in the area of interest for NIST (low false alarms), the VQG system matched the GMM based scheme.

The training time of our VQG system was 20% of the time taken with the standard E-M based GMM system. Note also that the GMM system used K-Means seeding to initialise the mixture means, weights and variances for an improved convergence rate. A fast mixture testing method was also implemented for both systems to improve the overall recognition speed. In this implementation, the closest five Gaussian mixture densities were accumulated for determining the output likelihood of an observation. The closest five mixtures were located using a fast city-block distance search. This process was faster than testing all mixture components with minimal performance loss.

6. Conclusion

This paper discussed the use of a VQG as a more efficient alternative to the standard Gaussian Mixture Model for relatively well-clustered data. The VQG was more robust to mismatched speaker recognition conditions for the multi-background speaker system. This was possibly attributed to the method of estimation of the VQG mixture means, weights and variances. The VQG method provides a rapid means of training and testing to form a reliable and efficient speaker verification system.

7. Acknowledgments

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

8. References

[1] D. Reynolds, “Speaker Identification and Verification using Gaussian Mixture Speaker Models”, *Speech Communication*, No 17, pp 91-108, 1995.

[2] Q. Lin, et al, “Selective use of the Speech Spectrum and a VQGMM Method for Speaker Identification”, *ICSLP*, pp 2415-2418, 1996.

[3] S. Slomka, “Multiple Classifier Structures for Automatic Speaker Recognition under Adverse Conditions”, *PhD Thesis*, Brisbane, Australia: Queensland University of Technology, 1999.

[4] A. Dempster, N. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *J. Royal Statistical Society*, Vol 39, pp 1-38, 1977.

[5] R. Schalkoff, *Pattern Recognition*, New York: John Wiley & Sons, 1989.

[6] D. Reynolds, “Comparison of Background Normalization Methods for Text-Independent Speaker Verification”, *Eurospeech*, Vol 2, pp 963-966, 1997.

[7] NIST’s Spoken Natural Language Processing Group Web Page: http://www.nist.gov/speech/.