ON THE CONVERGENCE OF GAUSSIAN MIXTURE MODELS: IMPROVEMENTS THROUGH VECTOR QUANTIZATION[1]

Moody, J., Slomka, S., Pelecanos, J. &, Sridharan, S.

Speech Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane, Q 4001, Australia
Email: jmoody@markov.sprc.qut.edu.au, s.sridharan@qut.edu.au
Fax: (+617) 3864 1516

ABSTRACT

This paper studies the reliance of a Gaussian Mixture Model (GMM) based closed-set Speaker Identification system on model convergence and describes methods to improve this convergence. It shows that the reason why the Vector Quantisation GMMs (VQGMMs) outperform a simple GMM is mainly due to decreasing the complexity of the data during training. In addition, it is shown that the VQGMM system is less computationally complex than the traditional GMM, yielding a system which is quicker to train and which gives higher performance. We also investigate four different VQ distance measures which can be used in the training of a VQGMM and compare their respective performances. It is found that the improvements gained by the VQGMM is only marginally dependant on the distance measure.

1. INTRODUCTION

Gaussian Mixture Models (GMMs) are used for a broad variety of statistical pattern recognition applications, including Speaker Identification (SI) [1]. It has been suggested that one factor which limits on the effectiveness of a Gaussian Mixture Model is its convergence while training [2]. The goal of this paper is to investigate and improve the convergence of the Model, and hence the accuracy, of the GMM in a closed-set SI system.

In [3] the authors have reported performance gains in the NIST evaluation training set when a GMM is preceded by a Vector Quantisation stage, yielding a VQGMM system. However, their paper neither investigates the reason for this improvement nor the application this may have to other systems.

In this paper, we perform a thorough investigation into this improvement in the VQGMM and conclude that it is due to improved convergence of the GMM. In addition, a relationship between clustering and model convergence is found and the complexity improvements over a normal GMM system investigated. Finally, a comparison of four VQ distance measures is made with respect to model convergence.

2. VQGMM SYSTEM

The VQGMM system consists of two stages; a Vector Quantisation algorithm and a Gaussian Mixture Model algorithm, as shown in Figure 1 [3].

[1]This project was supported by a research contract from the Defence, Science and Technology Office (DSTO).

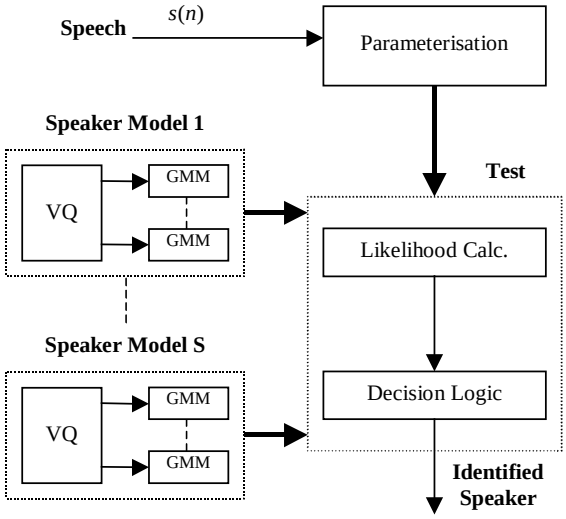


Figure 1: Block Diagram of the VQGMM System for SI

Initially, the data is broken into clusters using the k-means algorithm [4]. A multi-variate, continuous density GMM with nodal, diagonal covariance matrices [1] was trained on each cluster using the expectation-maximisation (EM) algorithm. As such, each speaker is characterised by a number of GMMs, trained on each VQ cluster.

When the system is used to recognise the speaker of a test utterance, the log-likelihood of each feature vector is calculated with respect to each individual GMM and each speaker. The value obtained is maximised over all clusters and speakers and the maximum value is labelled as the identified speaker. It should be noted that the testing phase involves no Vector Quantisation process.

Using this VQGMM technique, the authors of [3] were able to obtain a 10% reduction in error rates over the conventional GMM system.

It is argued in [3] that the model parameters can be better estimated by clustering the signal space into a number of smaller sub-spaces, such that feature vectors far away from the GMM have no effect on the GMM. However, no investigation has been carried out into these improvements in the estimation of model parameters.

3. DESCRIPTION OF THE SYSTEM

The VQGMM system was used to recognise forty-nine speakers in the wide-band portion of the King database [6]. The system was trained using the first session of the data and then tested on the remaining nine sessions.

Fifteen-dimensional Mel-Frequency FFT derived Cepstral Coefficients (MFCC) were used to parameterise the data. These coefficients were derived using a frame size of 32ms and a frame advance of 10ms. A low energy thresholding technique was used to remove silence from the speech.

In this study, the EM algorithm was run for a number of iterations, or until the change in training error of successive iterations (delta training error ΔE) fell below a certain threshold. Both the number of training iterations and the threshold were varied during testing.

During testing, the speech was divided into 5s segments with a 10ms frame advance. Each segment was then tested individually after the noise had been extracted.

4. VQGMM PERFORMANCE

Figure 2(a) and 2(b) show graphs of the performance of the VQGMM system, obtained with a number of VQ clusters (V) and GMM mixtures used for each speaker. Note that V=1 corresponds to the normal GMM without VQ.

The VQGMM system gives up to 5% improvement in recognition over a normal GMM. In addition, optimal performance is obtained with separation into V=15 VQ clusters with M=3 Gaussian Mixtures trained on each cluster.

5. VQGMM Convergence

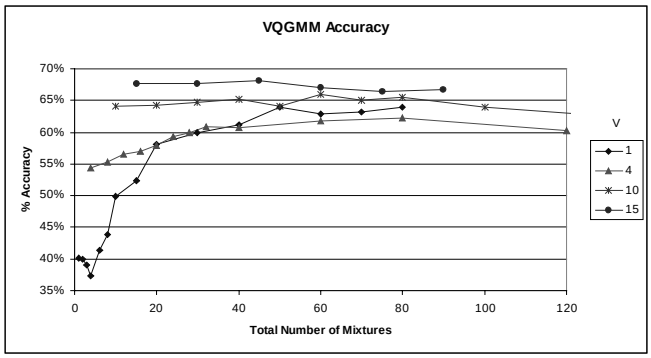
We also carried out a quantitative evaluation of the convergence of each model in the VQGMM system, by recording how many iterations of the EM algorithm are performed before GMM convergence. As mentioned above, this convergence criterion is determined by ΔE over each iteration of the EM algorithm.

This measure gives an indication of the complexity of the training data and the ease of training the GMMs on the data. If the change in training error fell below a certain threshold during the course of training, the model was said to have converged.

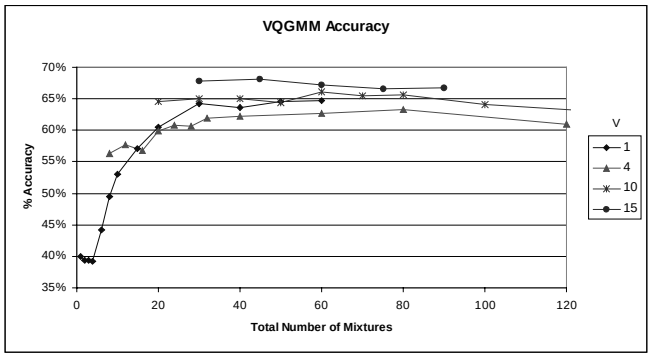
Figure 3(a) and 3(b) show the average number of iterations until model convergence for different VQGMM systems. This graph displays the total number of iterations for all 49 speakers.

It was found that the number of iterations taken for the GMM models in the optimal VQGMM systems was under 500, compared to 1800 for a normal GMM in the 49 speaker SI problem. It can be seen that the performance gains in each system investigated directly match convergence improvements.

These improvement gains over the original GMM system is partly due to the improvements in model convergence, as suggested in [3]. However, there are other reasons for this improved accuracy, such as the clustering itself assisting the GMM system. Clustering methods are investigated in Section 7.

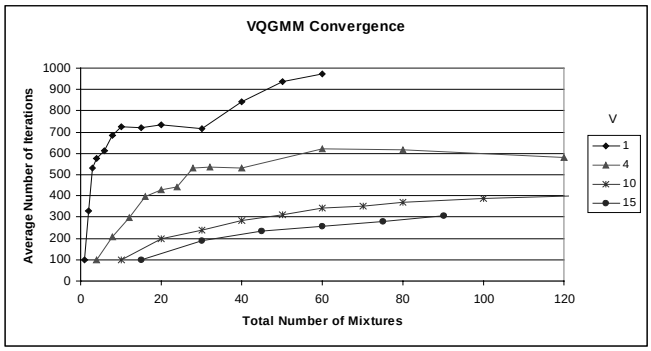


(a) 25 Iterations, Delta Training Error (ΔE) = 0.005

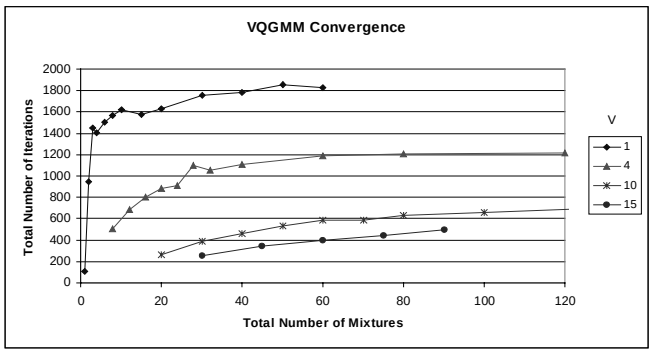


(b) 40 Iterations, Delta Training Error (ΔE) = 0.0005

Figure 2: Accuracy of the VQGMM System.



(a) 25 Iterations, Delta Training Error (ΔE) = 0.005



(b) 40 Iterations, Delta Training Error (ΔE) = 0.0005

Figure 3: Convergence of the VQGMM System.

6. COMPUTATIONAL EFFICIENCY

It can be shown [5] that the EM Algorithm for the training of the Gaussian Mixture Model can be broken into a number of stages. An investigation of these stages shows that the highest order complexity stage is the fifth stage when nodal, diagonal covariance matrices are used. This stage is involved with the mathematical evaluation of the equation:

$$SS_i^{pq} = \sum_{j=1}^{S} x_j^p x_j^q n_{ij} \qquad (1)$$

$$i = 1,2,...,M \text{ and } p = q = 1,2,...,D$$

where D is the dimensionality of the Feature Vectors, S is the number of sample vectors and M is the number of Gaussian mixtures used to model each cluster.

For I iterations of the EM Algorithm, this yields an order of computational complexity for the GMM system equal to

$$O_{GMM} = O(I \cdot D \cdot S \cdot M) \qquad (2)$$

as I, D, S and M become large.

In the VQGMM system using V clusters and the same number of mixtures overall, the order of computational complexity equation of the VQGMM and GMM may be compared using the following assumptions:

- The number of iterations (I) of the EM Algorithm is reduced by a factor C (~2), corresponding to the reduction in VQGMM iterations.
- The number of training Feature Vectors (S) for each GMM is reduced by a factor of V
- The number of Mixtures required in each GMM (M) is reduced by a factor of V
- The number of GMMs to be trained is increased by a factor of V
- The dimension of the feature vector D remains the same.

By combining these manipulations, the order of complexity of the VQGMM may be written as:

$$\begin{aligned} O_{VQGMM} &= O\left(I/C \cdot D \cdot S/V \cdot M/V \cdot V\right) \\ &= O\left(I \cdot D \cdot S \cdot M/C \cdot V\right) \end{aligned} \qquad (3)$$

As seen above, the amount of processing required to train the model is reduced by a factor equal to the product of the number of clusters by the improvement in convergence, for identical testing times. For the system above with 15 clusters and a convergence improvement factor of 2, this corresponds to a thirty-fold decrease in training time over the conventional GMM system.

It should be noted that this analysis does not take the computational complexity of the VQ clustering phase into account. An analysis of the VQ c-means algorithm yields the computational complexity to be of order $O(D \cdot S \cdot V)$, and does not greatly affect the VQGMM system.

7. CLUSTERING METHOD

We postulate that there are number of different reasons for the model convergence improvements using the VQGMM system.

The first reason is due to the system breaking the feature space into a number of homogenous sub-spaces. These sub-spaces are less complex and, as such, aid convergence of the model. The reduced complexity of the sub-spaces is apparent, due to the fewer number of mixtures required to model the sub-space.

A second reason for improved convergence is due to the fact that the data is clustered using first order moments (the cluster means). This may assist the GMM, which uses the first order moments as the centre of each mixture.

It was decided to investigate the inclusion of second order moments into the VQ clustering phase. The suitability of these clusters to future GMM classification can be measured, giving an insight into the features of a good clustering technique. In addition, this may also give an insight into the overall improvements attributed to the VQGMM system.

7.1. Clustering Methods

We compare the use of four systems to cluster the data and train the GMM, each using a different distance measure in the final VQ phase. The four methods use the Euclidean, the Weighted Euclidean, the Mahalanobis and the City Block distance measures respectively.

The four distance measures are defined below: [7]

Euclidean Distance: The Euclidean distance method accumulates the square difference between the two vectors

$$d_E(x, \bar{y}) = \sum_{i=1}^{N} (x_i - \bar{y}_i)^2$$

Weighted Euclidean Distance: The Weighted Euclidean distance method is identical to the Euclidean distance method but takes the reference vector variance components into account.

$$d_W(x, \bar{y}) = \sum_{i=1}^{N} (x - \bar{y})^T D^{-1} (x - \bar{y})$$

where D is the diagonal variance matrix.

Mahalanobis Distance: The Mahalanobis distance method includes the variance and correlation of the vector components.

$$d_M(x, \bar{y}) = \sum_{i=1}^{N} (x - \bar{y})^T W^{-1} (x - \bar{y})$$

where W is the square covariance matrix.

City Block Distance: The City Block distance, or absolute value distance, simply accumulates the absolute difference between each component of two vectors.

$$d_c(x, \bar{y}) = \sum_{i=1}^{N} |x_i - \bar{y}_i|$$

7.2. Results

We present results for the clustering methods using the four different distance measures. The system was tested using eight Vector Quantisation clusters with a number of GMM mixtures

in each cluster. The convergence criterion used was a Delta Training Error (ΔE) of 0.0005 and a maximum of 40 iterations of the EM algorithm were used.

The results of the investigation are shown in Figure 4.

As can be seen in the figure, there does not seem to be a great difference between the four different methods, although the Weighted Euclidean method does outperform the Euclidean method by around 1%. This is also reflected slightly in the convergence investigation, where the normal Euclidean method takes more iterations of the EM algorithm to converge. It is also interesting to note that any measure may have been used in the system, including the less computationally complex City Block distance measure.

This result adds more weight to the results obtained in the previous section; the improvements offered by the VQGMM system are due to the reduction in the model complexity for each GMM, assisting the EM algorithm.

8. CONCLUSION

This paper has performed an extended investigation into an inprovement to the GMM system; the VQGMM. It shows a relationship between the number of iterations required of the EM algorithm and improvements in accuracy and suggests that they may be due to improved EM Algorithm convergence.

A second component of this paper is a qualitative investigation into the computational complexity of the VQGMM system. It was found that the system offered an order improvement in model training complexity equal to the number of VQ clusters.

The final element of this paper was an investigation into the reliance of the VQGMM system on the distance measure used. Four distance measures were investigated, and it was found that there was little difference between them. It was found, however, that the normal Euclidean distance measure performed the most poorly and took the largest number of iterations to converge.

9. REFERENCES

1. Reynolds, A. and Rose, R. "Robust Text Independent Speaker Identification Using Gaussian Mixture Models", *IEEE Transactions on Speech and Audio Processing, Vol. 3, no. 1, 1995, pp72-83*.

2. del Alamo, C., Gil F., de la Torre Munilla, C. and Gomez, L. "Discriminative Training of GMM for Speaker Identification", Proceedings of *IEEE Conf. on Acoustics, Speech & Signal Processing, Vol 6, 1996, pp89-92*.

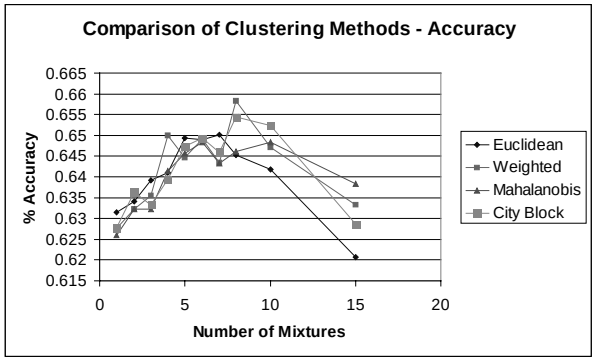
3. Qiguang, L., Jan, E., Che, C., Yuk, D. & Flanagan, J. "Selective Use of the Speech Spectrum and a VQGMM Method for Speaker Identification", *Proceedings of International Conf. on Spoken Language Processing, Vol 4, 1996, pp2415-2418*.

4. Schalkopf, R. *Pattern Recognition: Statistical, Structural and Neural Approaches*, McGraw-Hill, New York, 1997.

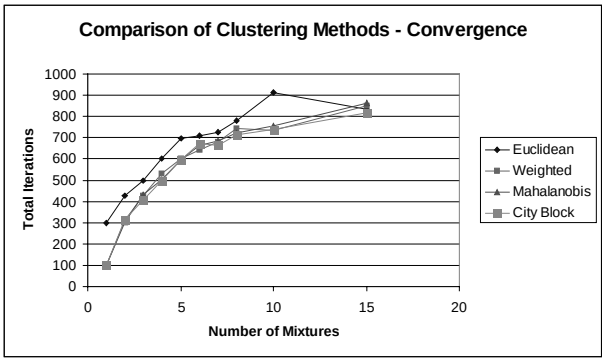
5. Zhang, Y., Alder, M. and Togneri, R. "Using Gaussian Mixture Modelling in Speech Recognition". *Proceedings of IEEE Conf. on Acoustics, Speech & Signal Processing Conference Paper, 1994, Vol I, pp613-616*.

6. Linguistic Data Consortium, *King Speech Corpus for Speaker Verification and Identification*, http://www.ldc.upenn.edu*/, 1992.*

7. Ong, S., Sridharan, S., Yang, C. and Moody, M. "Comparison of Four Distance Measures for Long Time Text-Independent Speaker Identification". Proceedings of *International Symposium on Signal Processing and its Applications, Vol 1, 1996, pp369-372*.



(a) Accuracy Results



(b) Convergence Results

Figure 4: Comparison of Clustering Methods