2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

Speaker Recognition and the Acoustic Speech Space

Robert Stapert and John S. Mason

Department of Electrical & Electronic Engineering,
University of Wales Swansea, SA2 8PP, UK

email: {eestaper, J.S.D.Mason}@swansea.ac.uk

Abstract

The hypothesis that for a given amount of training data a speaker model has an optimum number of components is examined. This is investigated with regard to Gaussian mixture models (GMM) with and without world model adaptation. Results show that maximising the number of components in a speaker model can improve speaker recognition results. Comparisons with vector quantisation (VQ) indicate that sensible use of out-of-class data is essential for optimising a recognition system.

1. Introduction

In [1] Dersch and King state that "the best model for a data set is the data set", i.e. that a set of training data is best modelled by itself. It follows that the corresponding best classifier is a simple nearest neighbour (NN), and in the limits this is true [2]. Dersch and King would seem to be extending this to all cases, not just when the data approaches infinity. This is an interesting hypothesis and one which might well be contended. An obvious caveat would centre on the use of out-of-class data.

In a Gaussian mixture model GMM each component consists of a mean, a covariance matrix and a weight. The density for component i of the model given the input vector $\vec{x}$ is given by Equation 1 where Σ_i is the covariance matrix and $\vec{\mu_i}$ is the mean vector. D is the dimension of the vector. A simplified form, popular in practical speaker recognition, has each component consisting of a mean, the diagonal of the covariance matrix and a weight.

$$b_i(\vec{x}) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_i|}} exp\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})'\Sigma_i^{-1}(\vec{x}-\vec{\mu_i})\} \quad (1)$$

In a GMM recognition test the probability of an input vector is given by Equation 2. Where $\vec{x}$ is the input vector and λ is the model; $b_i(\vec{x})$ is given in Equation 1, and w_i is the component weight; i ranges from $1 \ldots M$ where M is the number of components in the model. The probability $p(\vec{x}|\lambda)$ is, therefore, the weighted sum of M densities. In practice it is common to sum a small number of the most likely densities instead of all M, here the *five* most likely densities are summed.

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \quad (2)$$

A specific degeneration of this form to a minimal state leads to each component consisting of the mean, unity variance and a uniform weight. In the case where the likelihood of an input vector is given by the model component which is most similar to the input vector (instead of a weighted sum of all the components), we generally use the term vector quantisation (VQ) [3].

A GMM with a full covariance matrix is the most complex of the aforementioned models and VQ may be regarded as the least complex. Both GMM and VQ models require a measure of data reduction to calculate the components, i.e. from the no quantisation (NQ) extreme proposed in [1] where the model is the data, to a model of N components.

Assuming the use of in-class data only, then it can subsequently be hypothesised that for a set amount of data, the optimum number of model components is inversely linked to the complexity of the components, this is due to the greater amount of data required to estimate the more complex model components. Figure 1 illustrates the relationship between the complexity of a model and the amount of data required to accurately estimate the components. In the figure, NQ refers to no quantisation, VQ is vector quantisation, $GMMD$ is a GMM using the diagonal of the covariance matrix and $GMMF$ is the full Gaussian mixture model.

Here it is hypothesised that maximising the number of components in a model is important for speaker recognition, and in the limits, the number of components can equal the number of vectors in the training set. The bene-

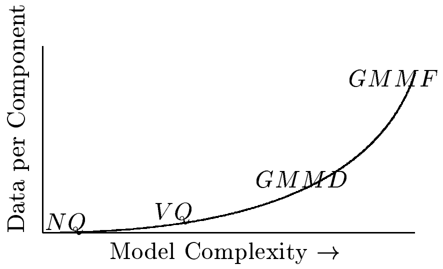

Figure 1: Illustration of the hypothetical NQ, VQ, GMM full ($GMMF$) and GMM diagonal ($GMMD$) training data requirements.

fit of using a larger model (more components) overrides the benefit of using more complex components [4]. This is in agreement with Dersch and King since the model with the largest number of components is the data itself, i.e. NQ. In Figure 2 the relationship between model complexity and the optimum number of components is illustrated.

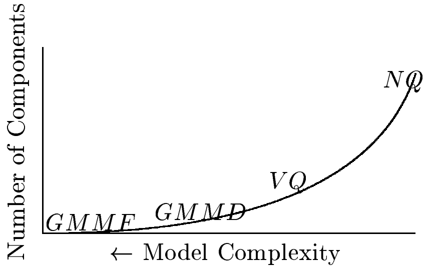

Figure 2: Illustration of the hypothetical optimum number of components for NQ, VQ, $GMMD$ and $GMMF$.

One implication of the above is that, for a given amount of training data, an optimised VQ (means only) should offer better speaker discrimination than an optimised GMM (means, variances and weights) if *no* out-of-class data is used. In an early work, Reynolds [5] compares VQ with two forms of GMM. The VQ and GMM models created for the experiments have approximately the same number of parameters (where a parameter is a single element of a component) and the speaker identification results are also approximately the same. It is suggested here that VQ is optimal when a larger number of parameters (and therefore a larger number of components) are used compared with GMM due to its less complex nature. Similarly, in [6] GMM and VQ models of the same size (same number of components) are compared and the GMM is found to be marginally better. Again, it is suggested here that a GMM model with N components may be optimal for a set amount of data, but a VQ model of the same data will be optimal with a larger number of components. This is a conclusion that is supported in [7] where the statement is made that the performance of VQ and GMM depends on the model size, and by Matsui in [8] where it is shown that VQ can outperform hidden Markov models if the number of VQ parameters is permitted to be optimised.

2. The Hypothesis

To improve the performance of a GMM, it is necessary for it to be able to approach or match the larger number of components that can be generated for VQ. This is where a world model [9] can be used. A world model (also often called a universal background model) is a model of the "world" of speakers. The world model makes it possible to train a GMM model with an equivalent number of components to that of VQ, thus maximising the use of speaker specific data. The number of GMM components that can accurately be estimated for a set amount of data is restricted due to the relatively large amount of data required to estimate statistics such as variances and weights. By adapting the speaker specific components using information taken from the world model this restriction is alleviated. Typically, the speaker data is modified according to a relevance factor W. An example is given in Equation 3 where R is a relevance constant for the world, a value of 16 has been found to be appropriate [10], and c is the number of speaker vectors in the cluster. The amount of data available for world model estimation is potentially unlimited, which makes it a valuable source of speech space statistics.

$$W = \frac{c}{c+R} \tag{3}$$

Speaker specific statistics are typically adapted according to Equation 4, where $\vec{\epsilon}_s$ is an element of the speaker model and $\vec{\epsilon}_w$ is the world model equivalent.

$$\vec{\epsilon}_s = W * \vec{\epsilon}_s + (1 - W) * \vec{\epsilon}_w \tag{4}$$

A VQ model can be viewed as having a uniform set of variances and weights, i.e. a special case of GMM; this is very unlikely to be optimum. It is arguably obvious that the same VQ model with accurate variances and weights will perform better. The logical extension is that GMM will always be better than VQ if it is known exactly *how* to utilise the world model statistics. And further, a full GMM will always be better than a diagonal GMM.

The hypothesis is illustrated for a given amount of data in Figure 3. The illustration has five hypothetical profiles labelled GMM, GMM_1, GMM_2, VQ_1 and VQ_2. GMM is an error curve for GMM's with increasing model sizes and without the aid of a world model. GMM_1 and GMM_2 use world model adaptation, the two curves differ in that they depict two possible trajectories relative to VQ. VQ_1 and VQ_2 are VQ error trajectories which use means-only and no world model. VQ_1 and VQ_2 differ in that they

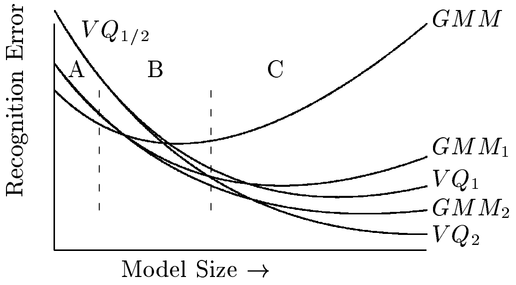

Figure 3: Hypothetical GMM and VQ performance curves.

illustrate two profiles relative to $GMM_{1/2}$. Two profiles are given to show the uncertainty of the individual merit of the two models.

Illustrated are the following predictions for increasing model size through sections A, B and C of the figure:

- Section A: Given enough speaker data, it is unnecessary and maybe detrimental for a GMM to use the world model. The amount of speaker specific data is sufficient for the estimation of accurate means variances and weights. These statistics are more representative of the speaker than any which could be gleaned from the world model (unless the world model is based on a near infinite amount of data); re-estimation using the world model is therefore ill-advised. Here the GMM profile shows less error than GMM_1/GMM_2 which use world model adaptation, and the GMM models generally outperform VQ_1/VQ_2 which are based on means only. In this section, because the model size is small, enough speaker specific data is available to calculate accurate speaker specific variances and weights for the GMM, and consequently it is the best.
- Section B: As the model size increases the error rates drop because more speech events are described. However, the training data is distributed over more components which gradually leads to less accurate GMM statistics, particularly variances and weights. Here, the world model may contain information which is more discriminative of the speaker, and it may help improve the speaker specific statistics by re-estimation. This is illustrated by GMM_1 and GMM_2 having lower error rates than GMM. The VQ_1 and VQ_2 error rates continue to drop because the means only components require less training data than the more complex GMM components, which permits VQ to benefit from having more components without the loss of accuracy in its statistics suffered shown by GMM.
- Section C: Here even less data is available for each component and the GMM statistics become increasingly inaccurate. Here, GMM_1 and GMM_2 diverge to illustrate two possibilities: given the high inaccuracy of the speaker specific statistics, adapting them with world model equivalents may not be the best solution and the recognition error rate may begin to increase, this is shown by the GMM_1 profile. An alternative approach is to *replace* the speaker specific variances and means with world model equivalents instead of re-estimating them. A model consisting of world model variances and weights and speaker specific means could continue to improve performance (as shown by GMM_2), since the GMM model will benefit from the increased number of components (similar to VQ) without the degree of accuracy loss demonstrated by the GMM and GMM_1 curves. It is the GMM_2 option which has proven to attain the better recognition rates [10].
- The VQ_1 and VQ_2 profiles diverge in section C of the figure to illustrate two possibilities with respect to GMM. They will both improve until just before the model is the data, i.e. NQ, after which (contrary to [1]) a minor increase in error is observed, as is shown in the experiments reported here. The question remains whether VQ will be superior to a GMM with world adaptation at this limit. VQ_1 has a higher recognition error rate than GMM_2 and VQ_2 a lower error rate to show the two possible trajectories. Many factors are likely to influence the outcome of this question, the most obvious perhaps are the quality of the world model and the method of adaptation.

Using the statistics from the world model obviously can be beneficial, and it has been found that replacing the speaker's variances and weights with those from the world model, rather than performing an adaptation, can improve performance [10]. This supports the GMM_2 hypothesis of Figure 3.

3. Experiments

The experiments presented here are designed to test the above hypothesis.

3.1. World Model Implementation

An initial world model estimate can be created by any suitable procedure. Here the procedure described in [3] and commonly known as LBG is used. The LBG procedure is adapted to provide variances and weights together with the means. This is similar to the extended vector quantisation (EVQ) described in [11] except that there complete covariance matrices are used whereas here only the diagonal is used. The initial model can then be passed to a likelihood maximisation algorithm which re-estimates the model components and increases the probability of the model given the data. Similarly to a speaker model, the size of the world model is limited because a suitable amount of data is required to estimate accurate statistics.

Here, a speaker model is created from the world model by adapting the world model towards the speaker data [10]. The speaker data are assigned to their nearest world model components. This creates a set of speaker data clusters where each centroid is linked to a component of the world model. For each cluster the mean is found and the centroid is adapted according to a relevance factor. The variances and weights are not changed. The speaker model thus consists of the original world model variances and weights and adapted means.

This implementation of the world model clusters the speaker data using the components of the world model as centroids. If the world model is large (a large number of components) in comparison with the amount of speaker data, the amount of speaker data clustered on each world model component in the training process can be small. This could result in a model with a large number of components, as for VQ, and taking the variances and weights from the world should add information to the means and give, as the result, a model with greater discriminating powers than VQ.

In testing the input speech is first compared with the world model and then with the speaker model, the difference between the similarity with the world and the similarity with the speaker is taken as the normalised measure.

3.2. Data

The data used is a subset of the BT Millar database. The BT Millar database is a 63 speaker multi-session digit database. For each of the 12 digits there are 25 versions. The versions were recorded in five sessions, each separated by approximately two weeks. The speech was clean, recorded using one microphone and sampled at 8KHz. The experiments use 20 speakers and 10 digits with 10 versions for training and 15 versions for testing, this gives 150 tests per speaker (3000 in total) for each recognition score. The world model is created from all the versions of a second set of 20 speakers. The feature vectors are 28'th order, 14 cepstra and 14 first-order regression.

3.3. Experimental Variables and Results

A GMM is tested on model sizes 1 to 1024 without the aid of a world model to verify that the recognition results initially improve as the number of components in the model increases, and afterwards degrade as the statistics (primarily the variances and weights) become less and less accurate. $GMM_{1/2}$ (GMM with world model adaptation) is tested using world model sizes 1 to 1024. In this case the size of the speaker model is proportional to the size of the world model. The recognition results are predicted to improve with increasing world model size since the number of speaker model components will increase as well. At a certain point the world model statistics will degrade and this will lead to an upturn in recognition error. VQ is tested with model sizes 1 to 1024. It is predicted that VQ will outperform GMM because it can benefit from the better results that go with larger models, without the detrimental effects of inaccurate variances and weights, however its performance relative to $GMM_{1/2}$ is in question. Finally a score is given for NQ, which is the model with the largest possible number of speaker specific components. The results are shown in Figure 4 and in Figure 5 from model size 16 onwards to focus on the important part of the profiles. The actual recognition scores can be read in Table 1.

Figure 5 shows the identification error rates for a GMM without a world model, $GMM_{1/2}$ with a world model and VQ which does not use out-of-class information. Comparing this figure to Figure 3 it is seen that the error profiles support the hypothetical profiles labeled GMM, GMM_2 and VQ_2. It is suggested in the hypothesis that GMM is superior to both VQ and $GMM_{1/2}$ when sufficient speaker specific data is present. This appears to be verified in Figure 5 where the speaker specific data is sufficient until model size 128.

GMM is best until model size 128, after this the amount of available speaker data is not sufficient to ensure accurate statistics and the results quickly degrade. Utilising world data allows $GMM_{1/2}$ to continue improving beyond model size 128 and reach an optimum at 512. At model size 1024, the world model statistics begin to degrade and the error increases once more. VQ continues to improve to model size 1024. The VQ model size is not limited by the need to calculate complex statistics and here can be twice the size of the best GMM model, this is seen to be beneficial in the fact that the VQ recognition error is lower than the $GMMN_{1/2}$ error. No quantisation (NQ) is plotted at the end of the VQ profile. It might be expected that this model will give the best results since it is the largest, however it is seen to produce an upturn in the error. This suggests that estimating a statistical model is beneficial since it could provide a better description of a speaker's speech space than the data itself (unless the data is infinite).

4. Conclusions and Discussion

The $GMM_{1/2}$ profile is limited by the amount of data used to create the world model. Given more data for the world model it would continue along the same trend as VQ. If enough out-of-class data is available, it should continue to improve beyond the optimum achieved by VQ. The fact

	1	2	4	8	16	32	64	128	256	512	1024	NQ
GMM	54.76	51.08	32.95	20.41	10.19	4.41	3.08	2.44	2.83	6.29	14.44	
GMM world	62.83	48.89	35.49	25.43	12.60	8.10	4.00	2.54	2.19	2.06	2.12	
VQ	62.00	57.67	42.53	30.27	17.17	8.80	4.43	3.07	2.30	2.03	1.90	3.67

Table 1: GMM, GMM plus world and VQ identification results for model sizes 1 to 1024 and NQ.

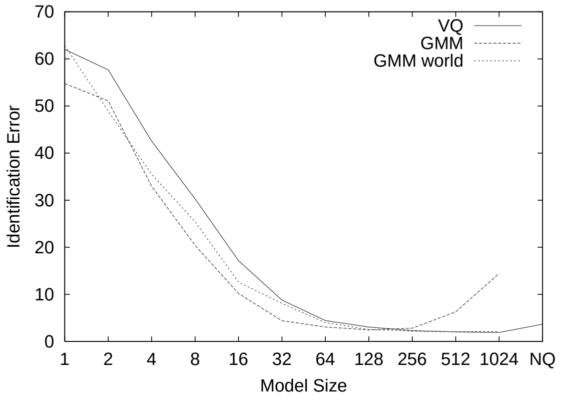

Figure 4: VQ, GMM without a world model and GMM with a world model are compared using model sizes 1 to 1024. The NQ score is appended to the VQ curve.

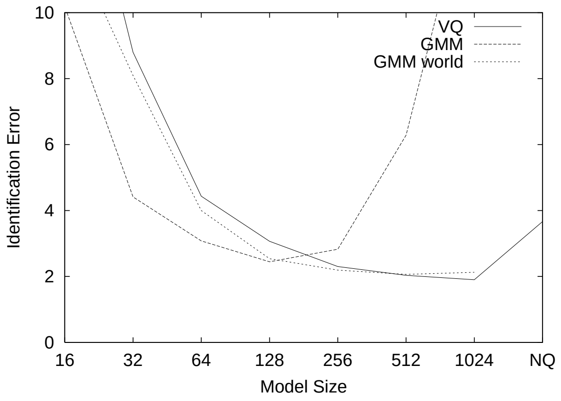

Figure 5: As for Figure 4 but from model size 16 onwards.

that NQ does not outperform VQ supports this argument as it points to the benefit of calculating statistics which describe the data.

It is concluded that maximising the use of speaker data, together with the sensible use of world statistics, is important to improve speaker recognition. Maximising the use of speaker data translates into maximising the size (number of components) of the model, i.e. VQ, or in the case of GMM, VQ plus information taken from the world of speakers. It might well be argued that GMM will always be the better speaker model if the best method to include information from the world is known and given a large quantity of world data. Further work will concentrate on optimising the utilisation of world data.

5. REFERENCES

1. D. R. Dersch and R. W. King. Speaker Models Designed from Complete Data Sets: A New Approach to Text-Independent Speaker Verfification. *Eurospeech97*, 5:2283, 1997.
2. A. L. Higgins, L. G. Bahler, and J. E. Porter. Voice Identification Using Nearest-Neighbor Distance Measure. In *Proc. ICASSP*, pages 375–378, 1993.
3. Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on COM*, 28:84–95, 1980.
4. R. P. Stapert. A Segmental Mixture Model, maximising data use with time sequence information. *Ph.D. Thesis, University of Wales Swansea*, 2000.
5. D. A. Reynolds and R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72 – 83, 1995.
6. S. van Vuuren. Comparison of Text-Independent Speaker Recognition Methods on Telephone Speech with Acoustic Mismatch. *ICSLP96*, 3:1788, 1996.
7. W. F. M. Chen and C. C. J. Kuo. Compressed Bit Stream Classification using VQ and GMM. In *SPIE Proceedings*, volume 3162, 1997.
8. T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proc. ICASSP*, volume 2, pages 157–160, 1992.
9. M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker verification system using alpha nets. In *Proc. ICASSP*, volume 1, pages 397–400, 1991.
10. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.
11. G. Kolano and P. Regel-Brietzmann. Combination of Vector Quantisation and Gaussian Mixture Models for Speaker Verification with Sparse Training Data. *Eurospeech 99*, 3:1203–1206, 1999.