ON THE USE OF DECOUPLED AND ADAPTED GAUSSIAN MIXTURE MODELS FOR OPEN-SET SPEAKER IDENTIFICATION

*J. Fortuna, A. Malegaonkar, A. Ariyaeeinia and P. Sivakumaran**

University of Hertfordshire, Hatfield, UK, *Canon Research Centre Europe Ltd., Bracknell, UK
{j.m.r.c.fortuna,a.m.ariyaeeinia,a.malegaonkar}@herts.ac.uk, *siva@cre.canon.co.uk

ABSTRACT

This paper presents a comparative analysis of the performance of decoupled and adapted Gaussian mixture models (GMMs) for open-set, text-independent speaker identification (OSTI-SI). The analysis is based on a set of experiments using an appropriate subset of the NIST-SRE 2003 database and various score normalisation methods. Based on the experimental results, it is concluded that the speaker identification performance is noticeably better with adapted-GMMs than with decoupled-GMMs. This difference in performance, however, appears to be of less significance in the second stage of OSTI-SI where the process involves classifying the test speakers as known or unknown speakers. In particular, when the score normalisation used in this stage is based on the unconstrained cohort approach, the two modelling techniques yield similar performance. The paper includes a detailed description of the experiments and discusses how the OSTI-SI performance is influenced by the characteristics of each of the two modelling techniques and the normalisation approaches adopted.

1. INTRODUCTION

Given a set of registered speakers and a sample utterance, open-set speaker identification is defined as a two stage problem [1]. Firstly, it is required to identify the speaker model in the set, which best matches the test utterance. Secondly, it must be determined whether the test utterance has actually been produced by the speaker associated with the best-matched model, or by some unknown speaker outside the registered set. The first stage is responsible for generating open-set identification error (OSIE). The decisions made in the second stage can generate either an open-set identification-false alarm (OSI-FA) or an open-set identification-false rejection (OSI-FR). This paper is concerned with open-set identification in the text-independent mode in which no constraint is imposed on the textual content of the utterances. It is well known that this is the most challenging class of speaker recognition. Open-set, text-independent speaker identification (OSTI-SI) is known to have a wide range of applications in such areas as document indexing and retrieval, surveillance, screening, and authorisation control in telecommunications and in smart environments.

One of the key issues in designing an OSTI-SI system is the selection of the type of speaker modelling technique. The Gaussian mixture model (GMM)-based approach is the most common choice for this purpose. In this technique, a speaker can be modelled by using either a decoupled-GMM [2] or an adapted-GMM [3]. In the former case, each model is built independently by applying the expectation maximisation (EM) algorithm to the training data from a specific speaker. In the latter case, each model is the result of adapting a general model, which represents a large population of speakers, to better represent the characteristics of the specific speaker being modelled. This general model is usually referred to as world model or universal background model (UBM). The common method used for the purpose of adaptation is based on the *maximum a posteriori* (MAP) estimation [4].

Two previous studies by the authors have independently investigated the performance of OSTI-SI using decoupled and adapted models respectively [1][5]. This paper combines the results obtained in the said previous studies and presents a comparative analysis on the use of decoupled and adapted Gaussian mixture models for the purpose of OSTI-SI.

The remainder of the paper is organised in the following manner. The next section describes the speech data, the feature representation and the GMM topologies used in the investigations. Section 3 details the testing procedure adopted, and Section 4 gives a summary of the score normalisations adopted. Section 5 provides a comparative analysis of the results obtained, and the overall conclusions are presented in Section 6.

2. EXPERIMENTAL CONDITIONS

The speech data adopted for this comparative study is based on a scheme developed for the purpose of evaluating OSTI-SI [1]. It consists of speech utterances extracted from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003. In total, the dataset includes 142 known speakers and 141 unknown speakers. The training data for each known speaker model consists of 2 minutes of speech and each test token from either population contains between 3 and 60 seconds of speech. These amount to a total of 5415 test tokens (2563 for known speakers and 2852 for unknown speakers). Achieving this number of test tokens is based on a data rotation approach which is detailed in [1]. For training the 2048 mixtures of the world model, all the speech material from 100 speakers is used (about 8 hours of speech). In the dataset there are also 505 development utterances from 33 speakers which can be used for score normalisation purposes.

In this study, each speech frame of 20ms duration is subjected to a pre-emphasis and is represented by a 16th order linear predictive coding-derived cepstral vector (LPCC) extracted at a rate of 10ms. The first derivative parameters are calculated over a span of seven frames and appended to the static features. The full vector is subsequently subjected to cepstral mean normalisation.

The GMM topologies used to represent each enrolled speaker in the studies involving decoupled and adapted models are 32m and 2048m respectively, where *N*m implies *N* Gaussian mixture densities parameterised with a mean vector and diagonal covariance matrices. In the case of the decoupled models, the parameters of each GMM are estimated using the maximum likelihood (ML) principle through a form of the expectation-maximisation (EM) algorithm [2]. In this case, an initial estimate of the model parameters for the EM algorithm is obtained by using a modified version of the LBG procedure, termed distortion driven cluster splitting (DDCS) [6]. In the case of the adapted models, the parameters of each GMM are estimated from the world model using a form of the MAP estimation procedure [3].

3. TESTING PROCEDURE

In each test trial, first, the following are obtained.

$$S_{\text{ML}} = \max_{1\le n\le N}\left\{\log\left(p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n)\right)\right\}, \tag{1}$$

$$n_{\text{ML}} = \arg\max_{1\le n\le N}\left\{\log\left(p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n)\right)\right\}, \tag{2}$$

If **O** is originated from the m^{th} registered speaker and $n_{\text{ML}} \neq m$ then an OSIE is registered and the score discarded. Otherwise, S_{ML} is normalised (with one of the score normalisation techniques considered in the next section) and stored in one of two groups depending on whether the observation is originated from a known or an unknown speaker. After the completion of all the test trials in a given investigation, the stored S_{ML} values are retrieved to form the empirical score distributions for both known and unknown speakers. These distributions are then used to determine the open-set identification equal error rate (OSI-EER), i.e. the probability of equal number of OSI-FA and OSI-FR.

When $\boldsymbol{\lambda}_n$ is a decoupled-GMM, the log-likelihood score for the sample utterance **O** as shown in (1) is computed as:

$$\log p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n) = \sum_{t=1}^{T}\left[\log\left(\sum_{c=1}^{N} w_c^{\lambda_n} b_c^{\lambda_n}(\mathbf{o}_t)\right)\right], \tag{3}$$

where $w_c^{\lambda_n} b_c^{\lambda_n}$ represents the weighted Gaussian probability density function for the c^{th} mixture in the n^{th} speaker model (or world model), *N* is the total number of mixtures in the speaker models and the world model respectively and *T* is the number of observations $\mathbf{o}_t$ in each test trial.

When $\boldsymbol{\lambda}_n$ is an adapted-GMM, the score is computed as:

$$\log\left(p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n)\right) = \sum_{t=1}^{T}\left[\log\left(\sum_{c=1}^{C} w_{\phi(c,t)}^{\lambda_n} b_{\phi(c,t)}^{\lambda_n}(\mathbf{o}_t)\right)\right], \tag{4}$$

where $w_{\phi(c,t)}^{\lambda_n} b_{\phi(c,t)}^{\lambda_n}$ represents the weighted Gaussian probability density function for the mixture given by $\phi(c,t)$ in the n^{th} speaker model (or in the world model). The function $\phi(c, t)$ represents the indexes of the *C* mixtures yielding the highest weighted probabilities for the feature vector $\mathbf{o}_t$ in the world model.

4. SCORE NORMALISATIONS

The scores computed according to equations (3) and (4) are affected by three main factors: distortions in the characteristics of the test utterance, misalignment of speaker models due to differences in the training conditions, and the problem of unseen data [3]. In order to tackle these problems, score normalisation methods can be used. The normalisations considered in this study are the world model normalisation (WMN), the cohort normalisation (CN), the unconstrained cohort normalisation (UCN), T-norm and various forms of Z-norm. Further details about these methods in the context of OSTI-SI can be found in [1].

5. EXPERIMENTAL RESULTS

Table 1 presents the results obtained for the considered modelling techniques in the first stage of the OSTI-SI. These results clearly show that the adapted-GMMs performed significantly better than the decoupled-GMMs. It appears that the coupling between the world model and each adapted-GMM seems to help the first stage of the OSTI-SI because of the better handling of the unseen data [3] as well as the contaminations of the test data.

	Decoupled-GMMs	Adapted-GMMs
OSIE (%)	33.7 ± 1.8	27.0 ± 1.7

Table 1: Relative performance of the considered modelling techniques in the first stage of OSTI-SI. The error rates are given with a 95 % confidence interval.

Table 2 shows the performance of the considered modelling techniques in the second stage of the OSTI-SI with various score normalisation methods. It also includes relative effectiveness of these modelling techniques without any form of score normalisation i.e. when the likelihood scores were determined according to equations (3) and (4).

Normalisation	Decoupled-GMMs	Adapted-GMMs
None	43.6 ± 2.4	47.8 ± 2.3
WMN	29.6 ± 2.2	22.9 ± 1.9
WMN_Z	26.8 ± 2.1	20.7 ± 1.8
CN	22.5 ± 2.0	20.7 ± 1.8
CN_Z	20.9 ± 1.9	19.1 ± 1.8
UCN	19.1 ± 1.9	18.5 ± 1.8
UCN_Z	20.7 ± 1.9	18.3 ± 1.8
T-norm	34.2 ± 2.3	18.6 ± 1.8
TZ-norm	29.6 ± 2.2	18.0 ± 1.7

Table 2: Results obtained in the second stage of the OSTI-SI (results are given in terms of OSI-EER(%) with a 95% confidence interval).

These results indicate that without any form of normalisation the use of adapted-GMMs leads to a higher error rate than that obtained with the decoupled-GMMs. This is thought to be due to the effect of the speaker independent components in each adapted-GMM. It should be noted that such an effect can be removed by using WMN and therefore it is common in the literature to consider the performance of the adapted-GMMs in conjunction with WMN as the baseline [3].

Table 2 shows that the adoption of WMN results in a significantly better result for the adapted-GMMs than for the decoupled-GMMs. Figure 1, which shows the DET curves obtained for WMN with these two modelling techniques, further confirms this relative effectiveness. At the first glance, it may be thought that this difference in performance is solely due to the better handling of the unseen data in the case of adapted-GMMs. However, it can be argued that in the second stage of OSTI-SI, this problem exists to a lesser extent. This is because a speaker model selected in the first stage is always the best match for the test utterance over all the registered speaker models. It is therefore felt that the difference in the observed performance is too significant for it to be solely attributed to the better handling of the unseen data by the adapted-GMM. It is thought that different GMM topologies for the speaker models and the world model could contribute to this difference.

It can be realised from Section 2 that, in the case of decoupled-GMMs, such a topological difference does exist. In this case, the speaker models are built with 32 mixtures whilst the world model consisted of 2048 mixtures. It is believed that with such a degree of topological difference, the contaminations in the test utterance could be reflected very differently in the best matched speaker model and the world model, compared to that in the case where the relevant models are of unique topology (which has been the case in adapted-GMMs). As a result, in the case of decoupled-GMMs, WMN may not be as effective as it is in the case of adapted-GMMs in compensating for such contaminations in the test utterance. In order to verify this hypothesis, a world model with 32 mixtures was trained using the same speech data as that for the 2048 mixture version. Table 3 presents the result of this study. It can be seen that, in this case, the performance of WMN improves significantly.

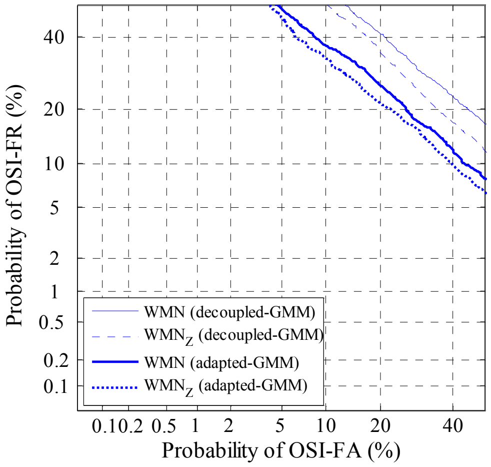

Figure 1: DET plots for the considered modelling techniques with WMN and WMN_Z

	World Model Topology	
	2048m	**32m**
OSI-EER (%)	29.6 ± 2.2	24.2 ± 2.0

Table 3: Effectiveness of the WMN for two different world model topologies (in the case of decoupled-GMM). (Results are given with a 95 % confidence interval).

Table 2, Figure 2 and Figure 3 indicate that in the cases of CN and UCN, the decoupled-GMMs offers similar levels of performance to those obtainable with the adapted-GMMs. It is also observed that the performance of the decoupled-GMMs followed that of the adapted-GMMs more closely in the case of UCN than in the case of CN. When the adapted-GMMs are used with CN/ UCN, the cohort speaker models have to take the role of handling the unseen data. These models cannot be as effective as the world model in accomplishing this task. This is because, in the case of CN and more in the case of UCN, there is no guarantee that the unseen data falls outside the adapted regions of the competing models. For the same reason, the performance obtained with CN and UCN in adapted-GMMs may not be considerably different from that in decoupled-GMMs. Based on the results obtained for CN and UCN, it appears that the cohort speaker models that are chosen based on their closeness to the best matched speaker model are better in accomplishing this task than the cohort speaker models chosen according to their closeness to the test utterance.

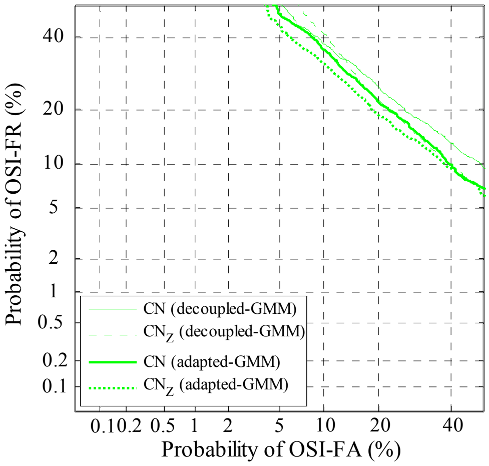

Figure 2: DET plots for the considered modelling techniques with CN and CN_Z.

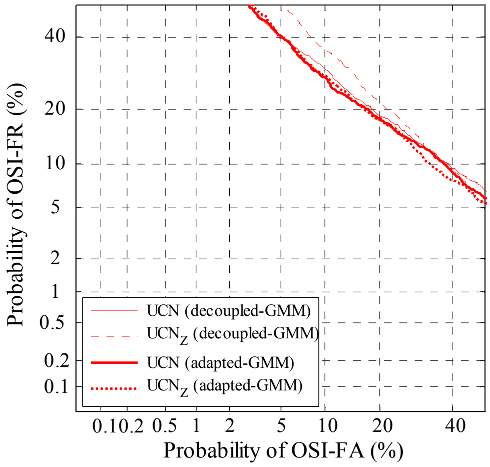

Figure 3: DET plots for the considered modelling techniques with UCN and UCN_Z.

It is interesting to note in Table 2 that the T-norm approach, which is one the worst performers in the case of the decoupled-GMMs, is one of the best performers in the case of adapted-GMMs. Figure 4 elaborates these results using DET curves. A careful investigation into these results shows that the reason for this is, to a large extent, dependent on how each registered speaker model reacts to a given test utterance produced by an unknown speaker. In the case of adapted-GMMs, this reaction is much similar across the registered speaker population, whereas in the case of decoupled-GMMs, it is considerably different. As a result, the T-norm parameters computed for adapted-GMMs tend to be much closer to those of the unknown speaker distribution and this makes the T-norm work better in this case. It should be noted that the Z-norm, which is specifically designed for aligning the models (i.e. reducing the model dependant biases), tend to produce more consistent reactions across the registered speaker population to a given test utterance produced by an unknown speaker. This may explain why, in the case of decoupled-GMMs, when T-norm is combined with Z-norm, a relatively large improvement is observed (Table 1 or Figure 4).

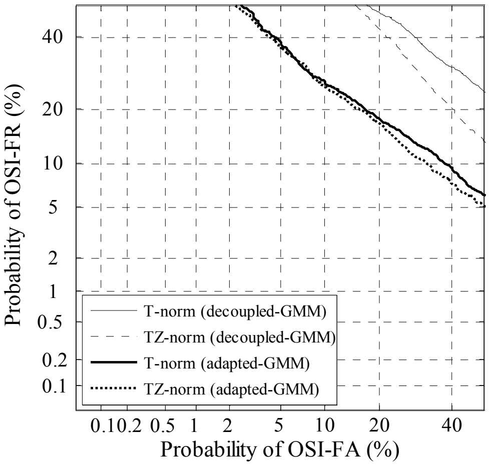

Figure 4: DET plots for the considered modelling techniques with T-norm and TZ-norm.

It is observed in Table 2 and Figures 1– 4 that with two exceptions, Z-norm exhibits similar levels of performance for both considered modelling techniques when it is combined with other normalisation methods. These exceptional cases are the T-norm and Z-norm combination (i.e. TZ-norm) which is discussed above, and the UCN and Z-norm combination (i.e. UCN_Z).

In the case of decoupled-GMMs, UCN_Z performs slightly worse than UCN. A close analysis of this case revealed that the underlying problem was the lack of availability of sufficient data for computing the Z-norm parameters for every known speaker model. In particular, it was observed that, with the available development data, the tail ends of the distributions assumed for computing the Z-norm parameters were significantly inaccurate. This problem may be tackled by adopting a large development set representing enough varieties of unknown speaker utterances. In other words, for each registered model, there should be an adequately large subset of the development data that can effectively be used as the unknown speaker utterances. Achieving this in practice is extremely difficult, especially when dealing with a large set of registered models. Therefore, it may be best to avoid the use of combined Z-norm and UCN with decoupled-GMMs. However, this problem is not as significant when decoupled-GMMs are replaced with adapted-GMMs. This is because, with adapted-GMMs, the scores produced by registered speakers for unknown utterances (in the development set) tend to be very similar. As a result, for each registered model, the validity of the Z-norm parameters obtained using the relevant subset of the development data is not too significantly influenced by the size of the subset. This may be the reason that, in the case of adapted-GMMs, UCN_Z does not achieve a worse error rate than UCN.

6. CONCLUSIONS

This paper has presented a comparative analysis of the performance of decoupled-GMM and adapted-GMM in OSTI-SI. It has been shown that, in general, the use of adapted-GMM results in better performance and this is particularly significant in the first stage of the OSTI-SI process. The better performance of the adapted-GMMs has been mainly attributed to the way in which such models handle the problem of the unseen data in the test segments. It was also found out that significant differences in the model topology limit the effectiveness of the WMN for the case of decoupled models. Furthermore, based on the experimental results it is shown that the cohort approaches are equally capable of achieving good performance with both types of models and this is found to be particularly evident for the case of UCN. It is also noted that T-norm is one of the worst performers in the case of decoupled-GMM despite being amongst the best performers in the case of adapted-GMM. Finally, the performance improvement achievable by Z-norm is similar with both modelling approaches with the exception of the cases involving UCN and T norm (i.e. UCN_Z and TZ-norm).

7. REFERENCES

[1] Fortuna, J., Sivakumaran, P., Ariyaeeinia, A. M., and Malegaonkar, A., "Relative Effectiveness of Score Normalisation Methods in Open-set Speaker Identification", Proc. Odyssey 2004 Speaker and Language Recognition Workshop, pp. 369-376, 2004.

[2] Reynolds, D., Rose, R. C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech Audio Proc., vol 3, 1995.

[3] Reynolds, D., Quatieri, T. F., and Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[4] Gauvain, J. L. and Lee, C.-H., "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. Speech Audio Process., vol. 2, pp. 291-298, 1994.

[5] Fortuna, J., Sivakumaran, P., Ariyaeeinia, A. M., and Malegaonkar, A., "Open-set Speaker Identification Using Adapted Gaussian Mixture Models", to appear in Proc. Interspeech'05, 2005.

[6] Ariyaeeinia, A. M. and Sivakumaran, P. "Comparison of VQ and DTW classifiers for speaker verification," Proceedings of the IEE European Convention on Security and Detection (ECOS'97), No. 437, pp. 142-146, April 1997.