Open-set Speaker Identification Using Adapted Gaussian Mixture Models

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

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 an investigation into the use of adapted Gaussian mixture models in the context of open-set, text-independent speaker identification (OSTI-SI). The study includes a scheme for using the fast-scoring method which has been proposed for speaker verification. Furthermore, it provides an evaluation of various score normalisation methods in the proposed OSTI-SI framework. The dataset used for the experimental investigation is based on NIST SRE2003 1-speaker detection task. It is shown that significant improvements can be achieved if only a single mixture is used in the fast-scoring technique. Furthermore, it is experimentally observed that comparable performance is obtained using unconstrained cohort normalisation, T-norm and TZ-norm. The paper provides a detailed description of the experimental set up, and presents an analysis of the results obtained.

1. Introduction

In general, speaker identification is the process of determining the correct speaker of a given test utterance from a registered population. If this process includes the option of declaring that the test utterance does not belong to any of the known speakers, then it is specifically referred to as open-set speaker identification. 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 utterance. 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, and constant authorisation control in systems involving man-machine dialogue in telecommunications.

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 [1] or an adapted-GMM [2]. In the former case, each model is built independently by subjecting the training data of a specific speaker to the expectation maximisation (EM) algorithm. 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 [3].

In practice, the training data of a specific speaker may not provide all the phonetic content of his/her test utterances. This can cause uncertainty in the evidence gathered in the test phase regarding the identity of the registered speakers. The adapted-GMMs provide a better mechanism to tackle this problem than that offered by decoupled-GMMs. This is because of their close coupling to the world model. It should also be noted that, due to the same reason, the adapted-GMMs have considerably larger footprints. Typically, an adapted-GMM consists of 2048 mixtures whereas the typical number of mixtures in a decoupled-GMM is 32. Having large footprints is a disadvantage in OSTI-SI, since all registered speakers have to be tested for each given test utterance. This becomes particularly significant when the registered population is large.

Based on practical observations, it has been proposed that the storage requirement for adapted-GMM can be reduced significantly [2]. Furthermore, in the context of speaker verification, a technique referred to as the fast scoring procedure is proposed for the adapted-GMM systems to speed-up the scoring process without any significant loss of accuracy [2]. It is shown that with this scoring method the adapted-GMM systems can surpass the operational speed of that of decoupled-GMMs. If this scoring method would have the same effect on OSTI-SI, there is no reason for not choosing adapted-GMMs for OSTI-SI over decoupled-GMMs. However, the nature of the OSTI-SI problem is somewhat different from that of speaker verification [4] and therefore, it is not possible to foresee the outcome of the effectiveness of this scoring method based on the results of speaker verification studies. Therefore, this paper sets out to study the impact of this scoring method in the OSTI-SI systems. The paper also investigates the way the score normalisation effectiveness in OSTI-SI is influenced by the said fast scoring method.

The remainder of the paper is organised in the following manner. The next section introduces the basic framework of the OSTI-SI problem. Section 3 provides a summary on building speaker models via MAP adaptation and details the specific implementation issues for the current task. Section 4 presents the score normalisation methods used in this study. Section 5 provides details of the speech data used and a description of the front-end processor. The experimental work, together with the results obtained is also discussed in this section. The overall conclusions are presented in Section 6.

2. Open-set speaker identification

Suppose that N speakers are enrolled in the system and their statistical model descriptions are $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_N$. If $\mathbf{O}$ denotes the feature vector sequence extracted from the test utterance, then the open-set identification can be stated as follows:

$$\max_{1\le n\le N}\{p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n)\} \gtrless \theta \;\rightarrow\; \mathbf{O} \in \begin{cases}\boldsymbol{\lambda}_i, i = \arg\max_{1\le n\le N}\{p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n)\}\\ \text{unknown speaker model}\end{cases} \quad (1)$$

where θ is a pre-determined threshold. Based on this equation, it is evident that open-set identification is a two-stage process. For a given **O**, the first stage determines the speaker model that yields the maximum likelihood, and the second stage makes the decision to assign **O** to the speaker model determined in the first stage or to declare it as originated from an unknown speaker.

It is possible that an observation **O**, which belongs to $\boldsymbol{\lambda}_m$, does not yield the maximum likelihood for $\boldsymbol{\lambda}_m$ in the first stage of the process. This would result in an error regardless of the decision made in the second stage. Even with a correct identification in the first stage, two types of errors are possible in the second stage: assigning **O** to one of the speaker models in the set when it does not belong to any of them, and declaring **O** which belongs to $\boldsymbol{\lambda}_m$, and yields the maximum likelihood for it, as originated from an unknown speaker. For the purpose of this paper, these three error types are referred to as ***OSIE***, ***OSI-FA*** and ***OSI-FR*** respectively (where OSI, E, FA and FR stand for open-set identification, error, false acceptance and false rejection respectively).

3. Adapted-GMMs approach

As noted in the introduction, in this approach, $\boldsymbol{\lambda}_n$, $n \in \{1\ldots N\}$, is generated by adapting a larger and well-trained speaker independent model (world model) using the training data from the n^{th} speaker. The adaptation process is accomplished by using a form of MAP estimation [3]. It consists of two distinct stages. In the first stage, the probabilistic alignment of each training feature vector with each of the mixtures in the world model is determined. This is subsequently used for the estimation of the sufficient statistics of the speaker's training data. In the second stage, the new mixture parameters derived from the sufficient statistics are combined with those of the world model using data-dependent mixing coefficients. These coefficients are used to control the proportion of the new and old estimates that remain in the final parameters. The criteria for controlling this proportion is based on the probabilistic alignment computed in the first stage in that higher probabilities allow a greater proportion of the new estimates to remain in the final parameters. This ensures that mixture parameters are adapted only if enough training data has been observed for the associated mixtures. On the other hand, mixture parameters which remain relatively unchanged are still derived from significant amounts of training speech observed during the estimation of the world model itself.

Due to the nature of the adaptation process, the resulting speaker specific models have larger footprints. This affects the operational speed of an open-set identification system significantly as the registered population becomes larger. In order to tackle this problem the fast scoring method proposed in [2] for speaker verification is considered in this study. This approach is based on two observed phenomena. The first is that when a large GMM is evaluated for a given feature vector, only a small number of mixtures contribute significantly to the likelihood value. The other phenomenon is that the components of the adapted GMMs retain direct correspondences with the mixtures of the world model so that feature vectors close to a particular mixture in the world model will also be close to the corresponding mixtures in the adapted speaker models. Using these two phenomena, the fast scoring technique for open-set identification is implemented as follows:

1. For each feature vector in the test segment $\mathbf{O}\equiv\{\mathbf{o}_1, \mathbf{o}_2,\ldots,\mathbf{o}_T\}$, determine the best C scoring mixtures in the world model.
2. Score each feature vector $\mathbf{o}_t$ against the corresponding C mixtures of each of the registered speaker models and compute $\log p(\mathbf{o}_t|\,\boldsymbol{\lambda}_n)$, $t \in \{1\ldots T\}$, and $n \in \{1\ldots N\}$.
3. Accumulate all the frame scores obtained in step 2 for each speaker model and select the model $\boldsymbol{\lambda}_n$ that yields the highest log-likelihood score.

Mathematically, the said scores are computed as:

$$\log\big(p(\mathbf{O}\,|\,\boldsymbol{\lambda}_n\big) = \sum_{t=1}^{T}\left[\log\left(\sum_{c=1}^{C} w^{\lambda_n}_{\phi(c,t)} b^{\lambda_n}_{\phi(c,t)}(\mathbf{o}_t)\right)\right] \quad (2)$$

where $w^{\lambda_n}_{\phi(c,t)} b^{\lambda_n}_{\phi(c,t)}$ represents the weighted Gaussian probability density function for the mixture given by $\phi(c,t)$ in the n^{th} speaker model (or world model). Effectively, the function $\phi(c,t)$ returns the indexes of the C mixtures (in the world model) that yield the highest weighted probabilities for the feature vector $\mathbf{o}_t$.

4. Score Normalisation

The scores computed according to equation (2) 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 which is mentioned in the introduction. In order to tackle this problem, score normalisation methods can be used. This section briefly describes the two major approaches to score normalisation considered in this study. Further details about these methods in the context of open-set identification can be found in [4]. The first approach, based on the Bayesian solution, involves the computation of a likelihood ratio which is usually given in the log domain as:

$$L(\mathbf{O}) = \log p(\mathbf{O}\,|\,\boldsymbol{\lambda}_{\text{ML}}) - \log p(\mathbf{O}\,|\,\boldsymbol{\lambda}_{\text{U}}) \quad (3)$$

This equation represents a relative log-likelihood score between the most likely speaker model $\boldsymbol{\lambda}_{\text{ML}}$ and $\boldsymbol{\lambda}_{\text{U}}$ which is the model representing the unknown speakers for the test token **O**. Since the model $\boldsymbol{\lambda}_{\text{U}}$ is unavailable in practice, in this study, the term $\log p(\mathbf{O}|\boldsymbol{\lambda}_{\text{U}})$ was approximated using the following techniques:

World model normalisation (WMN)

$$\log p(\mathbf{O}\,|\,\boldsymbol{\lambda}_{\text{U}}) \approx \sum_{t=1}^{T}\left[\log\left(\sum_{c=1}^{C} w^{\lambda_{WM}}_{\phi(c,t)} b^{\lambda_{WM}}_{\phi(c,t)}(\mathbf{o}_t)\right)\right], \quad (4)$$

where the subscript *WM* indicates the association of the world model and the remaining symbols have the same meanings as those defined for equation (2).

Unconstrained cohort normalisation (UCN)

$$\log p(\mathbf{O}\,|\,\boldsymbol{\lambda}_{\text{U}}) \approx \log\left(\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{O}\,|\,\boldsymbol{\lambda}_{\phi(k)})\right), \quad (5)$$

where, $\phi(i) \neq \phi(j)$ if $i \neq j$ and $\boldsymbol{\lambda}_{\phi}(1)$, $\boldsymbol{\lambda}_{\phi}(2)$, …, $\boldsymbol{\lambda}_{\phi}(K)$ are the models which yield the next K highest likelihood scores after

$\log p(\mathbf{O}|\boldsymbol{\lambda}_{\mathrm{ML}})$ which is the score obtained with the most likely speaker model.

The methods in the second approach aim to standardise one of the two score distributions associated with the known and unknown populations. However, due to practical issues, this is normally applied to the score distribution for unknown speakers. The normalisations in this approach have the following form:

$$L(\mathbf{O}) = \frac{\log\left(p(\boldsymbol{\lambda}_{\mathrm{ML}} \mid \mathbf{O}\right) - \mu_{\mathrm{U}}}{\sigma_{\mathrm{U}}} \tag{6}$$

where μ_{U} and σ_{U} are the estimated parameters for the unknown score distribution. However, it is not feasible to estimate the above parameters in the context of OSTI-SI [4]. For this reason, the methods in this approach are implemented as follows:

Zero normalisation (Z-norm)

$$L(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}_{\mathrm{ML}} \mid \mathbf{O}) - \mu_{\mathrm{Z}}(\boldsymbol{\lambda}_{\mathrm{ML}})}{\sigma_{\mathrm{Z}}(\boldsymbol{\lambda}_{\mathrm{ML}})}, \tag{7}$$

where $\mu_{\mathrm{Z}}(\boldsymbol{\lambda}_{\mathrm{ML}})$ and $\sigma_{\mathrm{Z}}(\boldsymbol{\lambda}_{\mathrm{ML}})$ are the mean and standard deviation of $\{\log p(\boldsymbol{\lambda}_{\mathrm{ML}}|\mathbf{O}_1), \log p(\boldsymbol{\lambda}_{\mathrm{ML}}|\mathbf{O}_2),\ldots, \log p(\boldsymbol{\lambda}_{\mathrm{ML}}|\mathbf{O}_I)\}$ and $\mathbf{O}_i \notin \boldsymbol{\lambda}_{\mathrm{ML}}$ is the i^{th} development utterance.

Test normalisation (T-norm)

$$L(\mathbf{O}) = \frac{\log p(\mathbf{O} \mid \boldsymbol{\lambda}_{\mathrm{ML}}) - \mu_{\mathrm{T}}(\mathbf{O})}{\sigma_{\mathrm{T}}(\mathbf{O})}, \tag{8}$$

where $\mu_{\mathrm{T}}(\mathbf{O})$ and $\sigma_{\mathrm{T}}(\mathbf{O})$ which are the mean and standard deviation of $\{\log p(\mathbf{O}|\boldsymbol{\lambda}_1), \log p(\mathbf{O}|\boldsymbol{\lambda}_2), \ldots, \log p(\mathbf{O}|\boldsymbol{\lambda}_{\mathrm{N}})\}$ and $\boldsymbol{\lambda}_n \neq \boldsymbol{\lambda}_{\mathrm{ML}}$ is the n^{th} speaker model.

It can be noticed that equation (7) involves an *a posteriori* probability. Since WNM, UCN or T-norm provide estimates of such probability, any of these normalisations can be used in conjunction with Z-norm. When WMN is used with Z-norm it is considered as the baseline, and simply referred to as Z-norm for the purpose of this paper. The term TZ-norm is used when Z-norm is used in conjunction with T-norm. In a previous study of OSTI-SI [4] it had been shown that the UCN could not be successfully combined with the Z-norm. For this reason this combination was not included in this paper.

5. Experimental investigation

5.1. Speech data

The speech data adopted for this study was based on a scheme developed for the purpose of OSTI-SI [4]. This dataset was put together with speech utterances extracted from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003. The dataset consisted of 142 known speakers and 141 unknown speakers. The training data for the known speaker models consisted of 2 minutes of speech and the test tokens for both populations contained between 3 and 60 seconds of speech. This accounted for a total of 5415 tests (2563 for known speakers and 2852 for unknown speakers). This number of test tokens was achieved through a data rotation approach which is detailed in [4].

For training the 2048 mixtures world model, all the speech material from 100 speakers was used (about 8 hours of speech). In the dataset there were also 505 development utterances from 33 speakers which were available for the score normalisation purposes.

5.2. Front-end processing

In this study, each speech frame of 20ms duration was subjected to pre-emphasis and was represented by a 16^{th} order linear predictive coding-derived cepstral vector (LPCC) extracted at a rate of 10ms. The first derivative parameters were calculated over a span of seven frames and appended to the static features. The cepstral mean was then subtracted from each of the feature vectors.

5.3. Testing procedure

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

$$S_{\mathrm{ML}} = \max_{1 \le n \le N} \left\{ \log\left(p(\mathbf{O} \mid \boldsymbol{\lambda}_n)\right) \right\}, \tag{9}$$

$$n_{\mathrm{ML}} = \arg\max_{1 \le n \le N} \left\{ \log\left(p(\mathbf{O} \mid \boldsymbol{\lambda}_n)\right) \right\}, \tag{10}$$

If $\mathbf{O}$ was originated from the m^{th} registered speaker and $n_{\mathrm{ML}} \neq m$ then an OSIE was registered and the score discarded. Otherwise, S_{ML} was normalised (with the considered score normalisation technique) and stored in one of two groups depending on whether the observation was 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 were retrieved to form the empirical score distributions for both known and unknown speakers. These distributions were 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.

5.4. Experimental results and discussions

The first experiment aimed at determining the speed/accuracy trade-off characteristic of the fast-scoring technique. It has been claimed in [2] that the frame likelihood values can be approximated very well using only the five best scoring mixtures. Figure 1 shows the error rates obtained for various numbers of best scoring mixtures.

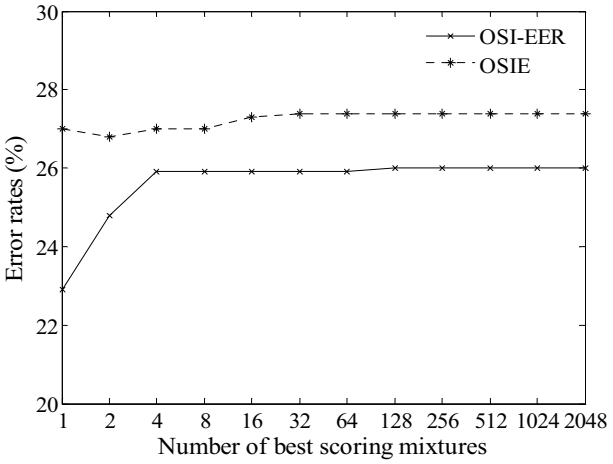

Figure 1: Effect of the number of best scoring mixtures on the OSIE and OSI-EER.

It appears that, in the case of OSTI-SI using only the best scoring mixture provides the highest capability for discriminating between known and unknown speakers. This is believed to be due to the selection of the best-matched models in the first stage of OSTI-SI. To be more specific, it has been observed that the unknown speakers who score very high tend to score high for a group of mixtures. As a result of this the overlap between the respective distributions of scores is expected to be considerably greater when using more than one mixture. It can also be seen in the figure that the performance seems to stabilise when more then four best scoring mixtures are used. This is in accordance with the observations made in [2]. Finally, it is also apparent from the figure that the OSIE is not significantly affected by the choice of the number of best scoring mixtures.

The next experiment aimed at evaluating the effect of score normalisation in the context of OSTI-SI with adapted models. In this case, based on the results obtained in the previous experiment, only the best scoring mixture was used. The results for the considered score normalisation methods are given in Table 1 and Figure 2.

Normalisation	OSI-EER ± CI95 (%)
WMN	22.9 ± 0.83
Z-norm	20.7 ± 0.80
T-norm	18.6 ± 0.77
UCN	18.5 ± 0.77
TZ-norm	18.2 ± 0.76

Table 1: Summary of the results.

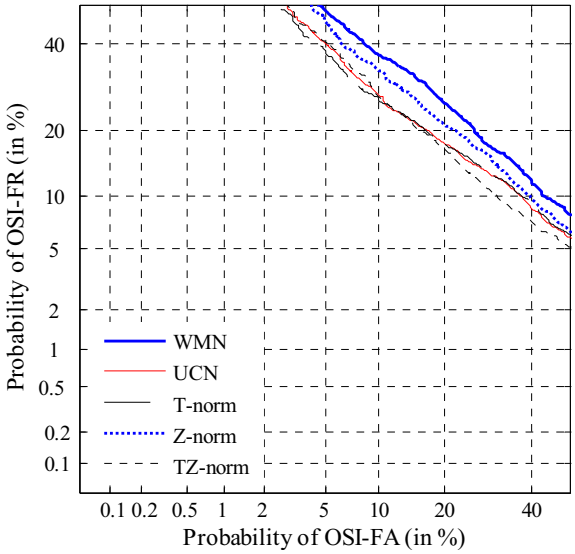

Figure 2: DET curves for various normalisation methods used in OSTI-SI.

It is clear from these results that, in general, score normalisation methods help to improve the performance of OSTI-SI. The use of Z-norm reduces the EER from about 23% to around 21%. This level of performance is further enhanced by using either of the T-Norm, TZ-norm or UCN which offer comparable levels of effectiveness. A study on the effectiveness of various score normalisations for OSTI-SI using decoupled models has been presented in [4]. In that study, it has been shown that T-norm is unable to obtain the performance improvement observed in the present experiments. Thus, it can be argued that the performance of T-norm is significantly influenced by the modelling technique adopted. To be more specific, it appears that the coupling amongst the adapted speaker models helps T-norm perform better. This can be attributed to smaller model-related variabilities in the adapted models compared with those observed in decoupled models. This is also supported by the relative improvement obtained with T-norm when the decoupled models have been pre-aligned using the Z-norm [4]. On the other hand the UCN seems to be less affected by the modelling approach adopted.

6. Conclusions

An implementation of OSTI-SI using adapted GMMs has been described. It has been shown that the large footprint associated with adapted models could be tackled using a form of a fast scoring technique previously proposed for speaker verification. Furthermore, it has been shown experimentally that, unlike in speaker verification, only the best-scoring mixture need to be included in the likelihood for achieving the best performance. Additionally, the study has confirmed the significance of score normalisation as a valuable component in OSTI-SI. It has been shown that, whilst Z-norm enhances the performance accuracy considerably, further improvement over the Z-norm performance can be achieved using either of UCN, T-norm or TZ-norm. Lastly, based on comparing the results in this study with those obtained earlier, it can be argued that, unlike the other normalisation methods considered, the effectiveness of T-norm is influenced by the modelling approach adopted, i.e. the method operates more effectively with adapted models.

7. References

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

[2] 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.

[3] 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.

[4] 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.

[5] Li, K. P., and Porter, J. E., "Normalisations and Selection of Speech Segments for Speaker Recognition Scoring", *Proc. ICASSP'88*, vol. 1, pp.595-598, 1988.

[6] Auckenthaler, R., Carey, M. and Harvey L. T, "Score Normalisation for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.