Verification effectiveness in open-set speaker identification

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

Abstract: Verification effectiveness in open-set, text-independent speaker identification is the authors' primary subject of concern. The study includes an analysis of the characteristics of this mode of speaker recognition and the potential causes of errors. The use of well-known score normalisation techniques for the purpose of enhancing the reliability of the process is described and their relative effectiveness is experimentally investigated. The experiments are based on the dataset proposed for the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003. On the basis of experimental results, it is demonstrated that significant benefits are achieved by using score normalisation in open-set identification, and that the level of this depends highly on the type of approach adopted. The results also show that better performance can be achieved by using the cohort normalisation methods. In particular, the unconstrained cohort method with a relatively small cohort size appears to outperform all other approaches.

1 Introduction

Speaker identification is a main subclass of automatic speaker recognition, defined as determining the correct speaker of a given test utterance from a registered population. When the process includes the option of declaring that the test utterance does not belong to any of the known (registered) speakers, then it is referred to as open-set speaker identification (OS-SI). A second subclass of speaker recognition is speaker verification (SV). This process involves determining whether a speaker is who (s)he claims to be. In this case, according to the degree of closeness of the test utterance to the target speaker model, a decision is made as to whether to accept or reject the claimant.

Given a set of registered speakers and a sample test utterance, the OS-SI process can be divided into two successive stages of identification and verification. Firstly, this is because it is required to identify the speaker model in the set that best matches the test utterance. Secondly, it must be determined (verified) 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 difficulty in this problem is exacerbated if speakers are not required to provide utterances of specific texts during identification trials. In this case, the process is referred to as open-set, text-independent speaker identification (OSTI-SI). This is the most challenging class of speaker recognition. It has a wide range of applications in such areas as document indexation, surveillance, and authorisation control in smart environments.

A factor influencing the complexity of OSTI-SI is the size of the population of registered speakers. As this population grows, the confusion in discriminating among the registered speaker voices increases. In addition, the growth in the said population also increases the difficulty in confidently declaring a test utterance as belonging to or not belonging to the initially nominated registered speaker.

The problem of OSTI-SI is further complicated by undesired variations in speech characteristics. These variations can have different causes ranging from environmental noise to uncharacteristic sounds generated by the speaker. The resultant variations in speech cause a mismatch between the corresponding test and pre-stored voice patterns from the same speaker. Such intra-speaker variations have been the subject of extensive study in recent years, mainly in the field of SV. The general problem in SV is that of minimising the overlapping between the score distributions for true speakers and impostors, so that it would be possible to verify or reject a claimed identity to a high degree of confidence using a preset threshold. The said mismatch between the testing and training material, however, has undesired effects on the score distribution parameters (i.e. variance and mean) for the true speaker. This can, in turn, lead to further overlapping of the score distributions for a true speaker and the impostors targeting that particular speaker. In practice, it is not possible to gather accurate information on the existence, level and nature of speech variations. In such cases, the most effective way to deal with this problem is score normalisation [1–7]. To date, a number of normalisation techniques have been developed, mainly with the aim of tackling the problem in the context of SV. In general, these techniques are based on either the Bayesian approach or the standardisation of the score distributions.

The problem in the second stage of OS-SI, however, is somewhat more challenging than that in the standard SV. This, which is further highlighted in Section 2, is due to the initial nomination of the speakers of the test utterances based on the best match-scores obtained in the first stage of

IEE Proceedings online no. 20050273

doi:10.1049/ip-vis:20050273

Paper first received 8th September 2005 and in revised form 27th February 2006

A.M. Ariyaeeinia, J. Fortuna, A. Malegaonkar are with the School of Electronic, Communication and Electrical Engineering, University of Hertfordshire, College Lane, Hatfield, Hertfordshire AL10 9AB, UK

P. Sivakumaran is with Canon Research Centre Europe Ltd, London Road, Bracknell, Berkshire RG12 2XH, UK

E-mail: a.m.ariyaeeinia@herts.ac.uk

the process. As a result, for example, each out-of-set speaker will have to be discriminated from the registered speaker who is its closest pair in the set. Because of the extended challenge in open-set identification and because of the differences in the characteristics of various score normalisation methods (Section 3), it may not be possible to foresee the effectiveness of the score normalisation methods in OS-SI from that obtained for SV. This is the focus of the investigations presented in this paper. It should be pointed out that there have previously been some studies on the use of score normalisation in speaker identification [8–11]. Some of these studies [8, 9] were concerned with the use of score normalisation at the sub-utterance (segmental/frame) level, which is not the subject of work in this paper. Moreover, in all the previous studies, only certain individual normalisation methods have been used for the benefit of closed-set identification. To date, the literature lacks a thorough evaluation of the relative effectiveness of various methods in the second stage of open-set identification.

2 Open-set speaker identification

Speaker identification involves representing a set of registered speakers using their corresponding statistical model descriptions, that is $\lambda_1, \lambda_2, \ldots, \lambda_N$, where N is the number of speakers in the set. Each model description is developed using the short-term spectral features extracted from the utterances produced by the registering speaker. On the basis of such speaker modelling, the process of speaker identification in the open-set mode can be stated as

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

where $\mathbf{O}$ denotes the feature vector sequence extracted from the test utterance and θ is a pre-determined threshold. In other words, $\mathbf{O}$ is assigned to the speaker model that yields the maximum likelihood over all other speaker models in the set, if this maximum likelihood score is greater than the threshold θ. Otherwise, it is declared as originated from an unknown speaker. On the basis of the above description, for a given θ, three types of errors are possible:

- $\mathbf{O}$, belonging to $\boldsymbol{\lambda}_m$, not yielding the maximum likelihood for $\boldsymbol{\lambda}_m$,
- assigning $\mathbf{O}$ to one of the models in the set when it does not belong to any of them and
- declaring $\mathbf{O}$ that belongs to $\boldsymbol{\lambda}_m$ and yields the maximum likelihood for it, as originated from an unknown speaker.

In this paper, these error types are referred to as OSI-E, OSI-FA and OSI-FR, respectively (where OSI, E, FA and FR stand for open-set identification, error, false acceptance and false rejection, respectively). Evidently, the first stage is responsible for generating OSI-E, whereas both OSI-FA and OSI-FR are the consequences of the decision made in the second stage.

It should be noted that an OSI-E in the first stage would always lead to an error regardless of the decision in the second stage. As the concern of this study is the second stage, the efforts should be on evaluating the verification reliability in the absence of any such identification errors in the first stage. In the experimental sense, this assumption involves discarding the false speaker nominations received from the first stage, when the actual speakers are within the registered set. Without such an assumption, a correct speaker rejection decision in the second stage would record a false rejection as far as the whole process is concerned.

In an OS-SI scenario, the universal speaker set is divided into two subsets of known (registered) speakers and unknown speakers. An important point to note is that each member of the unknown speakers can be falsely hypothesised as one of the registered speakers only (against whose model the unknown speaker achieves the highest score). In other words, for a fixed number of registered speakers, there are always a corresponding number of disjoint subsets of the unknown speakers. Each of these subsets contains the non-clients who all achieve their highest match-scores against one particular registered speaker. Any changes in the registered (known) speaker subset will result in corresponding changes in the number and membership of these non-client subsets. In practice, because of intersession variations, the membership of the said non-client subsets may not be entirely rigid, that is, an unknown speaker achieving its highest score against a particular model on one occasion, and against another registered model on a different occasion due, for example, to variation in his(her) speaking style.

To highlight the extent of difficulty in the second stage of open-set identification, the problem can be re-expressed as a special (but unlikely) scenario in the standard SV in which each impostor targets the speaker model in the system for which (s)he can achieve the highest score. This point is further illustrated in Fig. 1 which shows typical score distributions associated with these two forms of speaker recognition under the same experimental conditions. It should be pointed out that in the case of OS-SI, the client and non-client speakers are referred to as known and unknown speakers, respectively. In the case of SV, these are termed true and impostor speakers. As observed in this figure, the overlapping between the score distributions for unknown and known speakers in OS-SI is considerably greater than that between the score distributions for impostors and true speakers in SV. This is due to the selection of the best-matched models in the first stage of OS-SI, which has forced the score distribution mean for unknown speakers to be very close to that for the known speakers. It is also interesting to note that, for the same reason, the distribution variance for unknown speakers appears to be smaller than that for impostors in the case of SV.

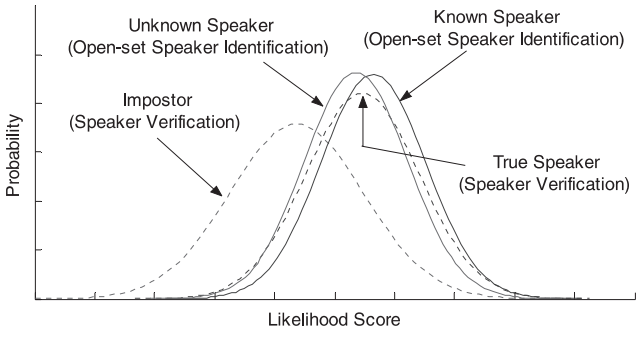

Fig. 1 *Score distributions associated with SV and the second stage of OSTI-SI*

It should be stated that the slight difference between the known and true speaker distributions is due to the fact that, in the case of OSTI-SI, the scores (associated with known speaker utterances) yielding an OSI-E are not included in the estimation of the known speaker distribution

3 Score normalisation

The main purpose of score normalisation is to help the separation between the score distributions for known and unknown speakers. In practice, this is particularly important because of the expected variations in speech characteristics. The effective reduction in the overlapping of the said distributions can lead to a reduction in OSI-FA and OSI-FR for a preset threshold. The following describes various methods in the two main categories of score normalisation highlighted earlier.

3.1 Bayesian solution

Under the Bayesian framework, the score required for making the decision in the second stage of open-set identification can be expressed as follows [5]

$$L(\mathbf{O}) = \log p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{ML}}) - \log p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{U}}) \tag{2}$$

where $\boldsymbol{\lambda}^{\mathrm{ML}} = \boldsymbol{\lambda}_i$, $i = \arg\max_{1 \le n \le N} \{p(\mathbf{O}|\boldsymbol{\lambda}_n)\}$, and $\boldsymbol{\lambda}^{\mathrm{U}}$ is the model representing the subset of unknown speakers that can falsely be hypothesised (in the first stage) as the speaker of $\boldsymbol{\lambda}^{\mathrm{ML}}$. In order to deploy (2), $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{U}})$ has to be determined accurately. However, in practice $\boldsymbol{\lambda}^{\mathrm{U}}$ is unavailable. Therefore the best option is to determine an appropriate replacement for $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{U}})$. For this purpose, the following three techniques can be adopted from the field of SV [1–7].

3.1.1 World model normalisation: This technique involves approximating $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{U}})$ with $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{WM}})$, where $\boldsymbol{\lambda}^{\mathrm{WM}}$ is a model generated using utterances from a very large population of speakers (such a model is commonly referred to as the world model (WM) [7] or the universal background model [6]).

It can be argued that the role of this normalisation method in OS-SI is to enhance the score for a known speaker when the test utterance is degraded. The assumption here is that both the reference model for the known speaker and the WM are free from all possible degradations (because of the use of clean training utterances or averaging-out the effects of contaminations in speech in the case of the WM). With such an assumption, it is not difficult to see that the existence of degradations in the test utterance will result in the scores against the known speaker model and the WM to be influenced in the same way (unfavourably). Consequently, the normalised score obtained using (2) should remain relatively unaffected.

The technique, however, does not aim to suppress the unknown speaker scores in relation to the scores for the corresponding known speakers. The reason is that the scores achieved by known and unknown speakers against a phonetically rich WM are, in general, very similar and any variations in these scores are not due to, or influenced by, the identity of the speakers.

3.1.2 Cohort normalisation: In this method, the model generated for each registered speaker is associated with a cohort of speaker models that are most competitive with it [2]. Here, the competitiveness of any two models is determined in terms of how close they are in the speaker space. The entire cohort selection is carried-out prior to the test phase, and $\log p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{U}})$ in (2) is approximated by

$$\rho_{\mathrm{CN}}(\mathbf{O}, \boldsymbol{\lambda}^{\mathrm{ML}}, K) = \frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{O}|\boldsymbol{\lambda}_{f(\boldsymbol{\lambda}^{\mathrm{ML}},k)}) \tag{3}$$

where $f(\boldsymbol{\lambda}^{\mathrm{ML}}, i) \neq f(\boldsymbol{\lambda}^{\mathrm{ML}}, j)$ if $i \neq j$ and $\boldsymbol{\lambda}_{f(\boldsymbol{\lambda}^{\mathrm{ML}},1)}$, $\boldsymbol{\lambda}_{f(\boldsymbol{\lambda}^{\mathrm{ML}},2)}$, ..., $\boldsymbol{\lambda}_{f(\boldsymbol{\lambda}^{\mathrm{ML}},K)}$ are the cohort of speaker models associated with $\boldsymbol{\lambda}^{\mathrm{ML}}$.

When the hypothesised speaker is a valid known speaker, the effect of score normalisation with this method should be similar in nature to that in the case of word model normalisation (WMN). In this respect, the approach in cohort normalisation (CN) may be viewed as deploying the most competitive subset of the WM speakers for each known speaker. As such, CN should be more effective than WMN in dealing with contamination in test utterances. This is because the cohort of models that are selected to be highly competitive with $\boldsymbol{\lambda}^{\mathrm{ML}}$ should provide a better replication of the way $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{ML}})$ is degraded by distortion in $\mathbf{O}$, than is possible with WM that is relatively diluted in terms of competitiveness. However, it should be pointed out that the dilution in normalisation factor obtained with WM is inherently limited. This is due to the fact that, in the generation of $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{WM}})$, there are relatively more significant contributions by WM subsets that are the closest to the observation $\mathbf{O}$, and therefore significantly more competitive with $\boldsymbol{\lambda}^{\mathrm{ML}}$ when $\mathbf{O}$ is produced by the true speaker.

In the use of CN in SV, it has already been found that, when the cohort size is very small, the normalisation procedure might potentially lead to the enhancement of impostor scores [5]. Considering a cohort size of one as an extreme case, it can be argued that the impostors, who score poorly against a target model, may be falsely accepted because of their scores against the single competing model being similarly low. In other words, the technique results in the enhancement of impostor scores relative to the true speaker scores. It has been shown that the above adverse effect of CN drops sharply as the cohort size is increased [5].

A similar behaviour of CN should be expected in OS-SI, with the exception that here the scores achieved by exclusive non-clients are normally relatively high (rather than low) against both their respective best matched-models and the corresponding competing models. As this is also the case when the hypothesised speakers are known-speakers, it appears that CN is not capable of unfavourably influencing the scores for unknown speakers relative to those for known speakers.

3.1.3 Unconstrained cohort normalisation: This method is similar to the cohort approach with the exception that the competing speaker models for each hypothesised speaker are selected during the test trial. To be more precise, $\log p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{U}})$ in (2) is replaced with

$$\rho_{\mathrm{UCN}}(\mathbf{O}, \boldsymbol{\lambda}^{\mathrm{ML}}, K) = \frac{1}{K}\sum_{k=1}^{K} \log p(\mathbf{O}|\boldsymbol{\lambda}_{\phi(k)}) \tag{4}$$

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 yielding the next K highest likelihood scores after $p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{ML}})$. Evidently, the method does not require any additional processing such as model generation/association prior to the test phase.

As indicated in (4), the competing speaker models are selected on the basis of their closeness to the test token. Therefore, in terms of enhancing a known speaker score, the unconstrained cohort normalisation (UCN) performance is at best similar to that of CN. This is because, due to the similarity of competing speaker models to the test token, the normalisation factor in this case is always greater than or at least equal to that in the case of CN. For the same reason, UCN provides a higher rate of suppression of scores for unknown speakers compared with WMN and CN. On the basis of (4), it is evident that the normalisation

factor in UCN is inversely related to the cohort size adopted. This gives rise to the question as to whether in OSTI-SI, the UCN cohort size can be determined such that it provides the best compensation when the test utterance from a known speaker is degraded (i.e. matching the CN performance), whereas still maintaining some effectiveness in terms of suppressing the scores for unknown speakers. Addressing this question is one of the aims of the experimental investigations in this paper. Indeed, the existence of such an optimum cohort size could enable UCN to outperform the other score normalisation procedures. On the other hand, because of variations in operating conditions, it might not be possible to determine one cohort size that is optimum in all experimental setups. However, the determination of some region of optimality could still be beneficial in unseen operating conditions.

3.2 Standardisation of score distributions

This approach was originally proposed for SV [12] with the aim of facilitating the use of a single threshold for all registered speakers. A major difficulty in setting a global threshold in SV is that both impostor score distribution and true speaker score distribution have different characteristics for different registered speakers. An approach to tackling this issue is that of fixing the characteristics of one of the score distribution types for all registered speakers. Currently, the common practice is to standardise the impostor score distributions. The main reason for operating on the impostor score distributions, rather than on the true speaker score distributions, is the unavailability of sufficient data (in the existing databases) for a reliable estimation of the standardisation parameters in the latter approach. As discussed subsequently, there are different approaches to computing the parameters for such a standardisation. In all cases, however, the computed parameters (i.e. mean and variance) are used for normalising the verification scores. The following presents the descriptions of the two main approaches in this category. The discussions are initially in the context of SV with the assumption that the impostor score distributions are Gaussian. This is then followed by a discussion on the deployment of the methods for OS-SI.

3.2.1 Zero normalisation (Z-norm): This method approaches the problem of score normalisation from the perspective of the speaker models. While aiming to standardise the impostor score distribution, the method provides an alignment of the registered speaker models, which are generated under different training conditions, prior to the test phase. In general, for each registered speaker model a single impostor distribution is obtained using a set of development impostor utterances. The mean and the standard deviation of the impostor distribution for each speaker model are then used for score normalisation as follows [7]

$$L_{\mathrm{SV}}(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}^{\mathrm{C}}|\mathbf{O}) - \mu_z(\boldsymbol{\lambda}^{\mathrm{C}})}{\sigma_z(\boldsymbol{\lambda}^{\mathrm{C}})} \tag{5}$$

where $\boldsymbol{\lambda}^{\mathrm{C}}$ is the model associated with the claimed identity (target speaker model), and $\mu_z(\cdot)$ and $\sigma_z(\cdot)$, which are specific to $\boldsymbol{\lambda}^{\mathrm{C}}$, represent the mean and standard deviation of the impostor score distribution. It is important to note that (5) involves a posteriori probability, suggesting that zero normalisation (Z-norm) should be used in conjunction with one of the score normalisation methods described in Section 3.1 or test normalisation (T-norm) that is discussed in Section 3.2.2. The reason is that, in order for this method to tackle any misalignments in the speaker models correctly, the adopted development impostor utterances should themselves be free from any misalignments. In practice, however, the development impostor utterances are misaligned because of various forms of speech anomalies. Therefore it is essential to enhance the alignment of these impostor utterances using another type of normalisation method before adopting them for Z-norm. In this case, (5) can be re-expressed to also reflect the use of score normalisation for impostor utterance alignment. The normalisation type adopted for this purpose must be consistent with that used in the subsequent test phase.

3.2.2 Test normalisation (T-norm): In this method, the required normalisation parameters are determined dynamically in the test phase using a set of example impostor models. The score normalisation in this case is based on the following equation [7]

$$L_{\mathrm{SV}}(\mathbf{O}) = \frac{\log p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{C}}) - \mu_T(\mathbf{O})}{\sigma_T(\mathbf{O})} \tag{6}$$

where $\mu_T(\mathbf{O})$ and $\sigma_T(\mathbf{O})$ are the mean and standard deviation of $\log p(\mathbf{O}|\boldsymbol{\lambda}_1^{\mathrm{EI}})$, $\log p(\mathbf{O}|\boldsymbol{\lambda}_2^{\mathrm{EI}})$, ..., $\log p(\boldsymbol{O}|\boldsymbol{\lambda}_J^{\mathrm{EI}})$ and $\boldsymbol{\lambda}_j^{\mathrm{EI}}$ is the jth example impostor model.

3.2.3 Deployment of Z-norm and T-norm in OSTI-SI: The direct adaptation of Z-norm and T-norm for open-set identification would result in the following two formulas

$$L(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}^{\mathrm{ML}}|\mathbf{O}) - \mu_Z(\boldsymbol{\lambda}^{\mathrm{ML}})}{\sigma_Z(\boldsymbol{\lambda}^{\mathrm{ML}})} \tag{7}$$

$$L(\mathbf{O}) = \frac{\log p(\mathbf{O}|\boldsymbol{\lambda}^{\mathrm{ML}}) - \mu_T(\mathbf{O})}{\sigma_T(\mathbf{O})} \tag{8}$$

where all the symbols have the same meanings as before except $\mu_T(\mathbf{O})$ and $\sigma_T(\mathbf{O})$ that are the mean and standard deviation of $\{\log p(\mathbf{O}|\boldsymbol{\lambda}_1), \log p(\mathbf{O}|\boldsymbol{\lambda}_2), \ldots, \log p(\mathbf{O}|\boldsymbol{\lambda}_L)\}$, with $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_L$ being the statistical models for L appropriately selected speakers. Ideally, $\boldsymbol{\lambda}_l$, $l = 1, 2, \ldots, L$, should be taken from a particular subset in the universal speaker set, whose members are the exclusive non-clients for $\boldsymbol{\lambda}^{\mathrm{ML}}$. In practice, it is not possible to follow this criterion. In fact, to avoid unnecessary computational costs, the registered speaker models are used for this purpose instead. A requirement in this case is that the set of registered speakers should be adequately large.

It is important to note that the above adapted versions of Z-norm and T-norm cannot lead to the standardisation of the score distribution for either any known speaker or any specific set of exclusive non-client speakers. In fact, what each of these methods attempts to achieve in OS-SI is to standardise the distribution of the general cross-speaker scores. This point is illustrated through the example presented in Fig. 2 for T-norm. It should be noted that similar processes in SV lead to the standardisation of the impostor score distribution.

4 Experimental investigations

4.1 Speaker representation

In all the experimental investigations, the speaker representation is based on Gaussian mixture models (GMM) [11]. The GMM topologies used to represent each enrolled speaker model and the WM are 32m and 2048m, respectively, where Nm implies N Gaussian mixture densities

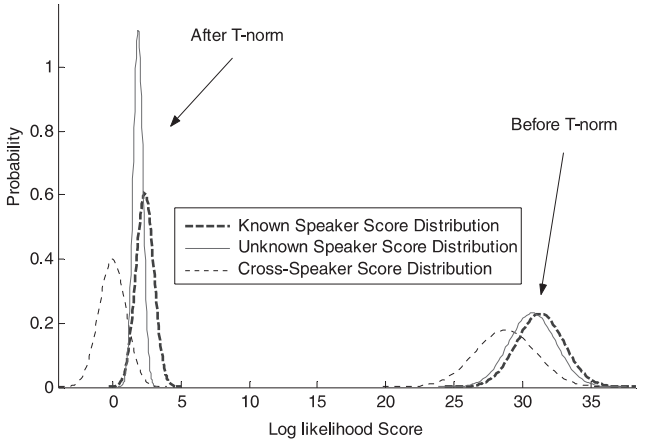

Fig. 2 *Typical plots of score distributions before and after applying T-norm*

each parameterised with a mean vector and diagonal matrix. The parameters of each GMM are estimated using a form of the expectation-maximisation algorithm [13].

The speech data adopted for the study is based on a scheme developed for the purpose of evaluating OSTI-SI [14]. It consists of speech utterances extracted from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003. The dataset includes 142 known speakers and 141 unknown speakers. The training data for each known speaker model consists of 2 min of speech and each test token from either population contains between 3 and 60 s 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 described in detail by Fortuna *et al.* [14]. The WM training is based on all the speech material from 100 speakers (about 8 h of speech). In the dataset, there are also 505 development utterances from 33 speakers that are used for the purpose of Z-norm.

In this study, the parametric representation of speech is as follows. Each speech frame of 20 ms duration is subjected to a pre-emphasis and is represented by a 16th-order linear predictive coding-derived cepstral vector extracted at a rate of 10 ms. 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.

4.2 Experimental results and discussions

Fig. 3 presents the equal error rates (EERs) in the second stage of OS-SI, obtained using different score normalisation methods. The figure also gives the EER without any score normalisation as the baseline. As observed, for the benefit of CN and UCN, these results are illustrated as a function of the cohort size. For the purpose of comparison, Fig. 4 presents the corresponding EERs obtained in SV experiments. An observation of these two figures clearing indicates the added difficulty in the case of OS-SI. This is reflected in the baseline EERs and is also observed in the results for various normalisation methods. Another immediate observation is the effectiveness of score normalisation methods in reducing EERs in both modes of speaker recognition.

Fig. 3 shows that the use of Z-norm in OS-SI results in the reduction of minimum achievable EER with all the normalisation methods except UCN (with relatively small cohort sizes). On the other hand, it is interesting to observe that

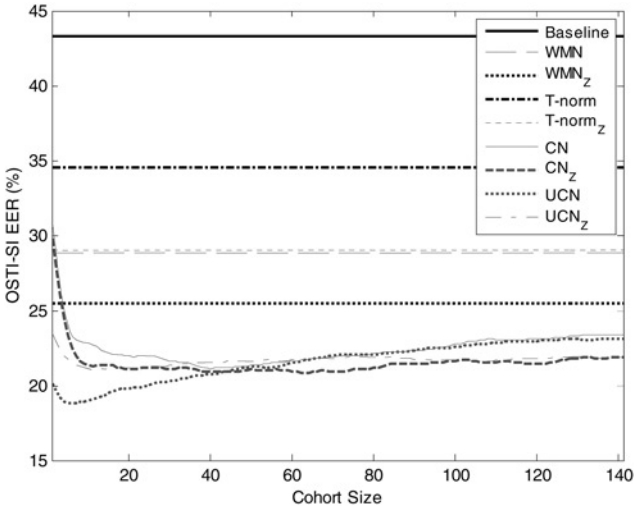

Fig. 3 *Comparison of the effectiveness of various normalisation methods in OSTI-SI in terms of EER*

the combination of Z-norm with UCN works well in SV (Fig. 4). This difference in effectiveness is believed to be due to the lack of availability of sufficient development data for computing the Z-norm parameters for every registered speaker model in OS-SI. To be more precise, in order for the combined Z-norm/UCN to have its maximum effectiveness, the UCN scenario in aligning the development utterances should match that in the case of test utterances from impostors/unknown speakers in the test phase. This is exactly the case in SV where the combined Z-norm/UCN works better than the UCN alone. It should be noted that in the test phase of SV, for each considered registered model, the impostor utterances achieve their highest scores mostly against other registered models in the set. This is highly similar to that happening when using the development utterances (from non-registered speakers) in extracting UCN-based Z-norm parameters. In contrast, in the test phase of open-set identification each registered model is targeted only by utterances from its own exclusive non-clients. Therefore by definition, the scores achieved by non-clients are always higher against the models they target than against any other registered models. This problem of scenario mismatch in OS-SI may be tackled by adopting a large development set representing enough varieties of unknown speaker utterances. In other words, for each registered

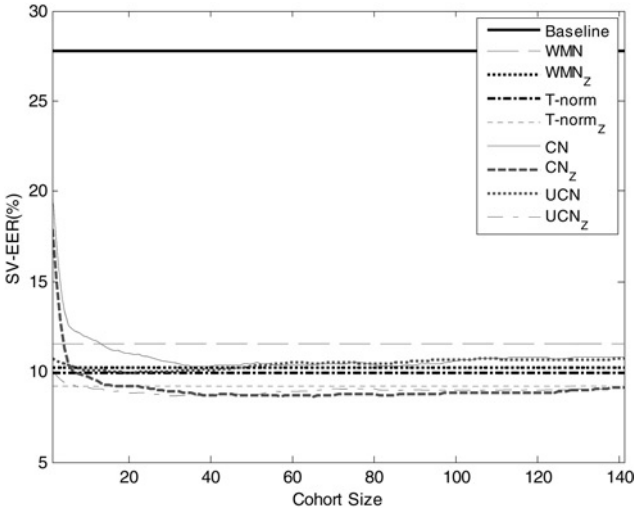

Fig. 4 *Comparison of the effectiveness of various normalisation methods in SV in terms of EER*

model, there should be an adequately large subset of the development data that can effectively be used as the utterances from exclusive non-clients (i.e. achieve their highest score against that particular registered model). 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/UCN with such a realisation of OS-SI. Finally, it is also worth noting that according to Fig. 4, the effects of the said scenario mismatch fades away for large cohort sizes (i.e. UCN_Z performs better than UCN). It is also observed that, for such cohort sizes, the EERs obtained with each of UCN and UCN_Z become very similar to those for CN and CN_Z, respectively. These are due to the fact that, for adequately large cohort sizes, UCN loses its unique property that differentiates it from CN, and is also the cause of the scenario mismatch in Z-norm.

As expected from the descriptions given in Section 3.1, CN performs rather poorly for small cohort sizes. It is also observed that (similar to the case in SV) the CN effectiveness in OS-SI improves sharply as the cohort size is increased.

The results in Fig. 3 show the UCN method as the best performer in OS-SI. In addition, it is observed that the minimum EER obtained with UCN is with cohort sizes of around 5–7. As suggested in Section 3.1, this appears to be the region of optimum cohort size for UCN. In other words, for cohort sizes in this region, UCN is effective in providing compensation when the test utterance from a known speaker is degraded while still maintaining capability in terms of suppressing the scores for unknown speakers.

Using the best cohort size for each of CN and UCN, Fig. 5 presents the experimental results in the second stage of OS-SI using the detection error trade-off (DET) curves. The plots clearly indicate the superior performance of cohort methods and, especially, UCN in open-set identification. Again, for the purpose of comparison, the corresponding DET curves obtained in SV experiments are illustrated in Fig. 6. A comparison of the results in these two figures (and also in Figs. 3 and 4) shows that T-norm, which is one of the best performers in SV (with and without Z-norm), provides the worst performance in OS-SI.

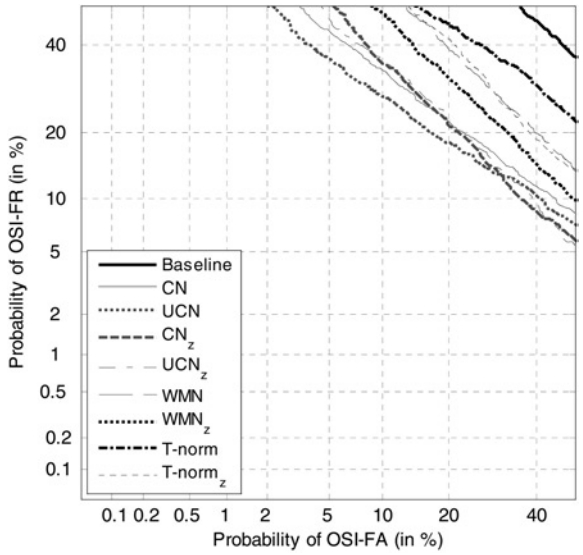

Fig. 5 *DET curves for various normalisation methods used in OS-SI*

The cohort sizes chosen for CN and UCN are those giving the best performance

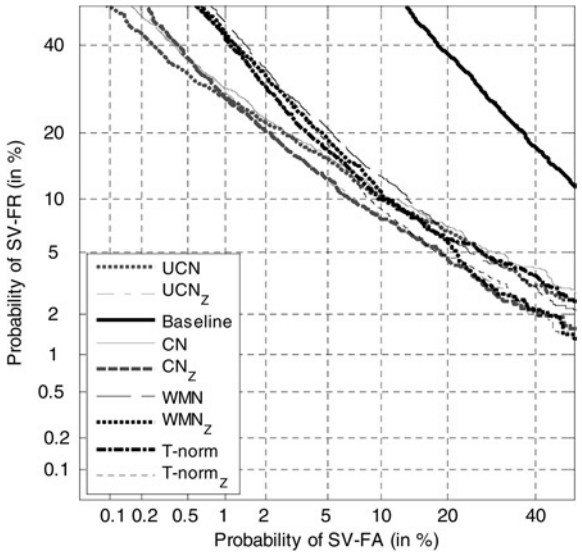

Fig. 6 *DET curves for various normalisation methods used in SV*

The cohort sizes chosen for CN and UCN are those giving the best performance

Similar to the case in SV, the WMN performance in OS-SI does not match that of UCN, or CN with an appropriately large cohort size. It is interesting to note that this difference in performance appears to be even wider in OS-SI. On the basis of the discussions in Section 3.1, the effectiveness of WMN relative to that of CN/UCN was not unexpected.

The best EERs obtained for individual score normalisation methods in OS-SI and SV are summarised in Table 1. The table also shows the relative improvement (RI) achieved using the normalisation methods over the baseline. The second and fourth columns in this table reiterate the fact that the EERs in OS-SI are, in general, larger than those in SV. It is also interesting to note that, moving from SV to OS-SI, there is a considerable drop in RI with every normalisation method (columns three and five in Table 1). Another immediate observation is that the incorporation of Z-norm always enhances the RI for the considered normalisation methods except in the case of UCN_Z in OS-SI.

According to Table 1, T-norm exhibits the sharpest drop (about 44%) in RI from SV to OS-SI. This result further highlights the reduction in the effectiveness of T-norm in OS-SI because of the inevitable compromise in its implementation (Section 3.2.3). The RI with UCN, on the other hand, is observed to sustain the lowest drop (about

Table 1: Results for the individual normalisation methods in terms of EER and RI (relative improvement)

Normalisation method	Best EER OS-SI (%) ± CI95	RI (%) OS-SI	Best EER SV (%) ± CI95	RI (%) SV
Baseline	43.3 ± 0.7	0	27.73 ± 0.07	0
T-norm	34.5 ± 0.6	20	9.92 ± 0.05	64
$T\text{-norm}_Z$	29.0 ± 0.6	33	9.24 ± 0.05	67
WMN	28.8 ± 0.6	34	11.57 ± 0.05	58
WMN_Z	25.5 ± 0.6	41	10.20 ± 0.05	63
CN	21.1 ± 0.6	51	10.29 ± 0.05	63
CN_Z	20.9 ± 0.6	52	8.65 ± 0.04	69
UCN	18.8 ± 0.5	57	9.96 ± 0.05	64
UCN_Z	21.1 ± 0.6	51	8.66 ± 0.04	69

7%) from the corresponding value in SV. This, together with the fact that UCN is the best performer in OS-SI, further confirms the fact that the added challenge in this mode of speaker recognition is one of dealing with the high match-scores by exclusive non-clients.

5 Conclusions

An investigation into the effectiveness of the verification process in the second stage of OS-SI has been presented. The study has provided valuable insight into certain important characteristics of this class of speaker recognition as well as into its performance features and limitations. It has been shown that an added challenge in the second stage of open-set identification, compared to standard SV, is due to the relatively high match-scores by unknown speakers. This problem is in addition to the difficulties caused by the mismatch (e.g. because of the contamination of speech) between the training and testing materials in practice. To minimise the adverse effects of these, the use of different score normalisation methods has been investigated. The outcomes have shown that, with or without normalisation techniques, the accuracy in the second stage of OSTI-SI is consistently below that in the standard SV. In addition, it has been found that in the case of OSTI-SI, the cohort methods exhibit the best performance.

The study has also shown that, because of practical limitations, the use of the standardisation methods in open-set identification can only lead to the standardisation of the general cross-speaker scores. However, as in OSTI-SI each registered model is targeted only by its own exclusive non-clients, it is concluded that (unlike in the case of SV) the standardisation methods cannot facilitate the use of a single decision threshold in this process.

An analysis of the performance of T-norm has shown that, in practice, this approach cannot be as effective in OSTI-SI as it is in SV. It has been found that, although T-norm is one of the top performers in SV, it provides the least relative improvement (about 20%) in OSTI-SI.

It has been shown that the use of Z-norm should be in combination with some other form of score normalisation to provide reliability in the model alignment process. The experimental results have confirmed that, except for UCN, the EERs obtainable with all other normalisation methods reduce noticeably when these are combined with Z-norm. The problem in the case of UCN_Z has been found to be due to the lack of availability of appropriately large varieties of utterances in the development dataset to meet the requirements in the OSTI-SI scenario. In addition, it has been shown that, in terms of reducing the scores for unknown speakers, UCN is less effective in OSTI-SI than is in SV. This is due to the selection of the best-matched model in the first stage of OSTI-SI, when the test material is produced by an unknown speaker. Nevertheless, UCN has been found to be the best performer in OSTI-SI, reducing the baseline EER by about 57%. This superior performance is believed to be in part because of the ability of the technique to exhibit some effectiveness in suppressing the scores for unknown speakers, while attempting to compensate for the adverse effects of contamination in test utterances from known speakers.

6 References

1 Higgins, A., Bahler, L., and Porter, J.: 'Speaker verification using randomized phrase prompting', *Digital Signal Process.*, 1991, **1**, (2), pp. 89–106

2 Rosenberg, A.E., DeLong, J., Lee, C.-H., Juang, B.-H., and Soong, F.K.: 'The use of cohort normalised scores for speaker verification'. Proc. Int. Conf. on Spoken Language Processing (ICSLP'92), 1992, vol. 1, pp. 599–602

3 Reynolds, D.A.: 'Speaker identification and verification using Gaussian mixture speaker models', *Speech Commun.*, 1995, **17**, (1–2), pp. 91–108

4 Rosenberg, A.E., and Parthasarathy, S.: 'Speaker background models for connected digit password speaker verification'. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96), 1996, vol. 1, pp. 81–84

5 Ariyaeeinia, A.M., and Sivakumaran, P.: 'Analysis and comparison of score normalisation methods for text-dependent speaker verification'. Proc. Eurospeech'97, 1997, pp. 1379–1382

6 Reynolds, D.A.: 'Comparison of background normalisation methods for text-independent speaker verification'. Proc. Eurospeech'97, 1997, pp. 963–966

7 Auckenthaler, R., Carey, M., and Lloyd-Thomas, H.: 'Score normalization for text-independent speaker verification systems', *Digital Signal Process.*, 2000, **10**, (1–3), pp. 42–54

8 Gish, H., and Schmidt, M.: 'Text-independent speaker identification', *IEEE Signal Process. Mag.*, 1994, **11**, (4), pp. 18–32

9 Markov, K.P., and Nakagawa, S.: 'Text-independent speaker recognition using non-linear frame likelihood transformation', *Speech Commun.*, 1998, **24**, (3), pp. 193–209

10 Rosenberg, A., Parthasarathy, S., Hirschberg, J., and Whittaker, S.: 'Foldering voicemail messages by caller using text independent speaker recognition'. Proc. Int. Conf. on Spoken Language Processing (ICSLP'00), 2000, vol. 2, pp. 474–478

11 Viswanathan, M., Beigi, H., Dharanipragada, S., Maali, F., and Tritchler, A.: 'Multimedia document retrieval using speech and speaker recognition', *Int. J. Doc. Anal. Recognit.*, 2000, **2**, (4), pp. 147–162

12 Li, K.-P., and Porter, J.E.: 'Normalizations and selection of speech segments for speaker recognition scoring'. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'88), 1988, vol. 1, pp. 595–598

13 Reynolds, D.A., and Rose, R.C.: 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (1), pp. 72–83

14 Fortuna, J., Sivakumaran, P., Ariyaeeinia, A., and Malegaonkar, A.: 'Relative effectiveness of score normalisation methods in open-set speaker identification'. Proc. Speaker and Language Recognition Workshop (Odyssey 2004), 2004, pp. 369–376