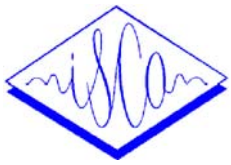

ODYSSEY04 - The Speaker and Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

RELATIVE EFFECTIVENESS OF SCORE NORMALISATION METHODS IN OPEN-SET SPEAKER IDENTIFICATION

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

University of Hertfordshire, College Lane, Hatfield, Hertfordshire, AL10 9AB, UK
*Canon Research Centre Europe Ltd, London Road, Bracknell, Berkshire, RG12 2XH, UK
*siva@cre.canon.co.uk, {j.m.r.c.fortuna, a.m.ariyaeeinia, a.malegaonkar}@herts.ac.uk

Abstract

This paper presents an investigation into the relative effectiveness of various well-known score normalisation methods in the context of open-set, text-independent speaker identification. The scope of the study includes a thorough experimental analysis of the performance of the methods considered. The experimental investigations are based on the use of the dataset proposed for the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003. The results clearly demonstrate that significant benefits can be achieved by using score normalisation in open-set identification, and that the level of this depends highly on the type of the approach adopted. Based on the experimental results, it is found that amongst the various normalisation methods considered, those which are based on the Bayesian solution provide the best performance. In particular, the unconstrained cohort method with a small cohort size appears to outperform all other approaches. The paper provides a detailed description of the experimental set up, and presents an analysis of the results obtained.

1. Introduction

Given a set of registered speakers and a sample utterance, open-set speaker identification is defined as a twofold problem. 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 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 and retrieval, surveillance, and constant authorisation control in systems based on ubiquitous computing and those involving man-machine dialogue in telecommunications.

The inherent complexity of OSTI-SI is dependent on the size of the population of registered speakers. As this population grows, the confusion in discriminating amongst the registered speakers' voices is likely to increase and therefore the number of incorrect identifications is likely to increase as well. The growth in the said population also increases the difficulty in confidently declaring a test utterance as not belonging to any of the registered speakers, when this is indeed the case. The reason is that, as the population size grows, the possibility of a voice originating from an unknown speaker being very close to one of the registered speaker models increases. It should, however, be noted that the analysis of the effects, on the OSTI-SI accuracy, of the registered population size is outside the scope of this paper.

The problem of OSTI-SI is further complicated by undesired variations in speech characteristics due to anomalous events. These anomalies can have different forms ranging from the communication channel and 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. This can in turn lead to degradation of the OSTI-SI performance. This paper is concerned with tackling this particular difficulty.

In practice, it is impossible to gather accurate information on the existence, level and nature of many speech distortions. In such cases, the most effective way to deal with this problem is known to be score normalisation [1-7]. This paper presents an analysis of various score normalisation methods for the purpose of OSTI-SI, and details a comparative evaluation of the effectiveness of these. It should be pointed out that the normalisation methods considered here have previously been investigated in the context of speaker verification (SV) [1-7]. However, the nature of the problem here is somewhat different from that of SV and therefore, it is not possible to foresee the outcome of this study from those of the SV studies.

It should further be pointed out that, although the use of certain score normalisation methods in open-set identification has previously been investigated [8-11], the literature lacks a thorough experimental evaluation of the relative effectiveness of the methods that can be adopted for this purpose.

The paper is organised in the following manner. The next section looks at open-set identification from a mathematical perspective. The considered score normalisation methods are detailed in Section 3. Section 4 describes the protocol used for the experimental investigation and provides details on the adopted speaker modelling methods. The experimental work

together with the results obtained are also discussed in this section. The overall conclusions are presented in Section 5.

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:

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

where 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 system, if this maximum likelihood score itself is greater than the threshold θ. Otherwise, it is declared as originated from an unknown speaker. It is evident from the above description that, for a given θ, three types of error are possible:

- $\mathbf{O}$, which belongs to $\boldsymbol{\lambda}_m$, not yielding the maximum likelihood for $\boldsymbol{\lambda}_m$.
- Assigning $\mathbf{O}$ to one of the speaker models in the system when it does not belong to any of them.
- Declaring $\mathbf{O}$ 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 types of error 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).

Based on equation (1), it is evident that open-set identification is a two-stage process. For a given $\mathbf{O}$, the first stage determines the speaker model that yields the maximum likelihood, and the second stage makes the decision to assign $\mathbf{O}$ to the speaker model determined in the first stage or to declare it as originated from an unknown speaker. Of course, the first stage is responsible for generating OSIE whereas, both OSI-FA and OSI-FR are the consequences of the decision made in the second stage.

An important point to note in this two-stage process is that the latter stage is far more susceptible to distortions in the characteristics of the test utterance than the former stage. This is because, in the former stage, since the same test utterance is used to compute all the likelihood scores, the distortions in the test utterance are likely to be similarly reflected in all the likelihood scores. As a consequence, the selection of the model that yields the maximum likelihood is likely to be unaffected. On the other hand, in the second stage, the absolute maximum likelihood score is compared against a threshold determined a priori and without any knowledge about the characteristics of the distortion in the test utterance. This inherent difficulty in the second stage is the primary focus of this paper.

It should be pointed out that a task similar to that described above (in the second stage of open-set identification) is also encountered in speaker verification. However, in this case, the problem is not as challenging. To be more specific, the challenge in open-set identification can be viewed as a special (but unlikely) scenario in speaker verification in which each impostor targets the speaker model in the system for which he/she can achieve the highest score.

This point is further illustrated by Figure 1 which shows typical score distributions associated with these two forms of speaker recognition under the same experimental condition. As observed, the overlapping between the score distributions for unknown and known speakers in open-set identification is considerably greater than that between the score distributions for impostors and true speakers in speaker verification.

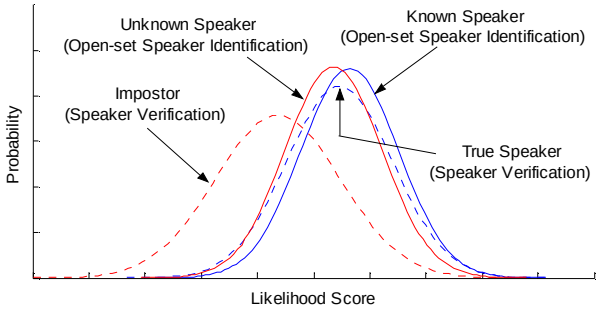

Figure 1: Score distributions associated with SV and the second stage of OSTI-SI (it should be stated that the slight difference between the known and true speaker distributions is due to the fact that, in the case of OSTI-SI, those scores associated with known speaker utterances which yield an OSIE are not included in the estimation of the known speaker distribution).

3. Score normalisation

3.1. Bayesian Solution

The decision rule for the second stage of open-set identification can, in general, be expressed as follows.

$$P(\lambda^{\mathrm{ML}}\mid\mathbf{O}) \gtrless P(\lambda^{\mathrm{U}}\mid\mathbf{O}) \rightarrow \mathbf{O}\in\begin{cases}\lambda^{\mathrm{ML}}\\ \lambda^{\mathrm{U}}\end{cases}, \quad (2)$$

where $\lambda^{\mathrm{ML}}=\boldsymbol{\lambda}_i, i=\arg\max_{1\le n\le N}\{p(\mathbf{O}\mid\boldsymbol{\lambda}_n)\}$, and λ^{U} is the model representing the unknown speakers. By applying the Bayes' theorem to the inequality in (2), it can be shown that

$$\frac{p(\mathbf{O}\mid\lambda^{\mathrm{ML}})}{p(\mathbf{O}\mid\lambda^{\mathrm{U}})} \gtrless \frac{P(\lambda^{\mathrm{U}})}{P(\lambda^{\mathrm{ML}})} \rightarrow \mathbf{O}\in\begin{cases}\lambda^{\mathrm{ML}}\\ \lambda^{\mathrm{U}}\end{cases}, \quad (3)$$

where $p(\mathbf{O}|\lambda^{ML})/p(\mathbf{O}|\lambda^{U})$ is the score to be computed in this stage and $P(\lambda^{U})/P(\lambda^{ML})$ is the threshold that has to be determined a priori. In practice, a more convenient form of representing the above score is

$$L(\mathbf{O}) = \log p(\mathbf{O}|\lambda^{ML}) - \log p(\mathbf{O}|\lambda^{U}) . \quad (4)$$

In order to realise the benefit of this Bayesian solution fully, $p(\mathbf{O}|\boldsymbol{\lambda}^{U})$ has to be determined accurately. In practice, however, $\boldsymbol{\lambda}^{U}$ (which represents the model for unknown speakers) is unavailable. Therefore, the best option is to determine an appropriate replacement for $p(\mathbf{O}|\boldsymbol{\lambda}^{U})$ so that, at least, some of the benefits of the resulting Bayesian solution can be retained. A situation similar to this also occurs in speaker verification. In that case, the Bayesian solution yields the following score [2][5].

$$L_{SV}(\mathbf{O}) = \log p(\mathbf{O}|\lambda^{C}) - \log p(\mathbf{O}|\lambda^{I}) , \quad (5)$$

where $\boldsymbol{\lambda}^{C}$ is the model associated with the claimed identity and, $\boldsymbol{\lambda}^{I}$ is the impostor model which is, in fact, unavailable in practice. Various techniques have already been proposed in order to tackle this problem in speaker verification [1-7]. Based on these techniques, three methods can be derived to deal with the problem described above for open-set speaker identification. These methods are as follows.

3.1.1. World Model Normalisation (WMN)

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

3.1.2. Cohort Normalisation (CN)

In this method, the model generated for each registered speaker is associated with a cohort of speaker models which are found to be most competitive with it [2]. Here, the competitiveness of any two speaker 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 $p(\mathbf{O}|\boldsymbol{\lambda}^{U})$ for a cohort of K speakers is computed as:

$$\rho_{CN}(\mathbf{O}, \lambda^{ML}, K) = \log\left(\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{O}|\lambda_{f(\lambda^{ML},k)}) \right), \quad (6)$$

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

3.1.3. Unconstraint Cohort Normalisation (*UCN*)

Unlike the previous two methods, this method does not require any additional processing such as model generation/association prior to the test phase. Here, log p($\mathbf{O}|\boldsymbol{\lambda}^{U}$) is replaced with

$$\rho_{UCN}(\mathbf{O}, \lambda^{ML}, K) = \log\left(\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{O}|\lambda_{\phi(k)}) \right), \quad (7)$$

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 p($\mathbf{O}|\boldsymbol{\lambda}^{ML}$). This method can also be viewed as a special case of the CN method where the required cohort of speakers is chosen according to the closeness of speaker models to the test utterance (this concept has been introduced by the authors in [5]).

3.2. Standardisation of Score Distributions

The methods in this approach, like the techniques described in the previous section, were originally proposed for speaker verification [6][7]. In their original form, they aim to standardise each form of the impostor score distribution, resulting from a different operating condition. The reason for operating on the impostor score distribution, rather than on the true speaker score distribution, is to obtain more reliable estimates for the standardisation parameters. Further details of the two main methods in this category are given below, initially in the context of speaker verification, and with the assumption that the impostor score distributions are Gaussian. This description is then followed by a discussion on the adaptation of the methods for open-set speaker identification.

3.2.1. Zero Normalisation (Z-norm)

This method is unique in the sense that it approaches the problem of score normalisation from the perspective of the speaker models. It is primarily concerned with the mismatches in the training conditions (e.g. use of carbon-button and electret microphones). In fact, the aim of Z-norm is to align the speaker models, which are generated under different training conditions, prior to the test phase.

There are two forms of Z-norm. In the first form, each registered speaker model is associated with a set of impostor score distributions. The parameters of these distributions (i.e. the mean and the standard deviation pairs) are computed using different sets of development impostor utterances. Each of these sets is associated with a different operating condition. Here, the score normalisation is performed according to the following equation:

$$L_{SV}(\mathbf{O}) = \frac{\log p(\lambda^{C}|\mathbf{O}) - \mu_{z}(\lambda^{C}, \varphi(\mathbf{O}))}{\sigma_{z}(\lambda^{C}, \varphi(\mathbf{O}))} , \quad (8)$$

where $\boldsymbol{\lambda}^{C}$ is the target speaker model, $\mu_z(\cdot)$ and $\sigma_z(\cdot)$ are specific to $\boldsymbol{\lambda}^{C}$ and represent the mean and standard deviation of the impostor score distribution for the operating condition given by $\varphi(\mathbf{O})$. In this case, a scheme has to be devised to detect the operating condition in the test phase to invoke the correct set of parameters $\mu_z(\cdot)$ and $\sigma_z(\cdot)$. Since the operating condition needs to be known explicitly, this form of Z-norm

is well suited only to certain practical applications [6]. This form is not considered in the present experimental work due to the nature of the database used.

In the second form, the operating condition does not need to be known explicitly. In this case, each registered speaker model is associated with a single impostor score distribution. The parameters of this distribution are computed using a set of development impostor utterances. Here the Z-norm is based on the following equation [7]:

$$L_{\mathrm{SV}}(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}^{\mathrm{C}} \mid \mathbf{O}) - \boldsymbol{\mu}_{\mathrm{z}}(\boldsymbol{\lambda}^{\mathrm{C}})}{\boldsymbol{\sigma}_{\mathrm{z}}(\boldsymbol{\lambda}^{\mathrm{C}})}, \tag{9}$$

where the symbols have similar meanings as those in equation (8). Unlike the first form, this form focuses purely on alleviating misalignment in the speaker models. In other words, it does not account for the speech characteristic distortions in **O** in anyway.

It can be noticed that equations (8) and (9) involve *a posteriori* probabilities. This implies that Z-norm should be used in conjunction with the score normalisation methods described in Section 3.1 or the T-norm method (which is discussed in the next section). For example, in the case of Z-norm with WMN (Section 3.1.1), equation (9) would have the following form:

$$L_{\mathrm{SV}}(\mathbf{O}) = \frac{\log \dfrac{p(\mathbf{O} \mid \boldsymbol{\lambda}^{\mathrm{C}})}{p(\mathbf{O} \mid \boldsymbol{\lambda}^{\mathrm{WM}})} - \boldsymbol{\mu}_{\mathrm{z}}^{\mathrm{WM}}(\boldsymbol{\lambda}^{\mathrm{C}})}{\boldsymbol{\sigma}_{\mathrm{z}}^{\mathrm{WM}}}, \tag{10}$$

where

$$\boldsymbol{\mu}_{\mathrm{z}}^{\mathrm{WM}}(\boldsymbol{\lambda}^{\mathrm{C}}) = \frac{1}{I}\sum_{i=1}^{I} \log \frac{p(\mathbf{O}_i^{\mathrm{DIU}} \mid \boldsymbol{\lambda}^{\mathrm{C}})}{p(\mathbf{O}_i^{\mathrm{DIU}} \mid \boldsymbol{\lambda}^{\mathrm{WM}})}, \tag{11}$$

$$\boldsymbol{\sigma}_{\mathrm{z}}^{\mathrm{WM}}(\boldsymbol{\lambda}^{\mathrm{C}}) = \frac{1}{I-1}\sqrt{\sum_{i=1}^{I}\left(\log \frac{p(\mathbf{O}_i^{\mathrm{DIU}} \mid \boldsymbol{\lambda}^{\mathrm{C}})}{p(\mathbf{O}_i^{\mathrm{DIU}} \mid \boldsymbol{\lambda}^{\mathrm{WM}})} - \boldsymbol{\mu}_{\mathrm{z}}^{\mathrm{WM}}(\boldsymbol{\lambda}^{\mathrm{C}})\right)^2}, \tag{12}$$

and $\mathbf{O}_1^{\mathrm{DIU}}, \mathbf{O}_2^{\mathrm{DIU}}, ..., \mathbf{O}_I^{\mathrm{DIU}}$ are the development impostor utterances.

3.2.2. Test Normalisation (T-norm)

In this method, the required transformation parameters are determined dynamically in the test phase by using a set of example impostor models. Like Z-norm, it has two forms. The first form does not require any explicit knowledge of the operating condition and it is based on the following equation [7]:

$$L_{\mathrm{SV}}(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}^{\mathrm{C}} \mid \mathbf{O}) - \boldsymbol{\mu}_{\mathrm{T}}(\mathbf{O})}{\boldsymbol{\sigma}_{\mathrm{T}}(\mathbf{O})}, \tag{13}$$

where $\boldsymbol{\mu}_{\mathrm{T}}(\mathbf{O})$ and $\boldsymbol{\sigma}_{\mathrm{T}}(\mathbf{O})$ are the mean and standard deviation of $\log p(\boldsymbol{\lambda}_1^{\mathrm{EI}} \mid \mathbf{O})$, $\log p(\boldsymbol{\lambda}_2^{\mathrm{EI}} \mid \mathbf{O})$, …, $\log p(\boldsymbol{\lambda}_J^{\mathrm{EI}} \mid \mathbf{O})$ and $\boldsymbol{\lambda}_j^{\mathrm{EI}}$ is the j^{th} example impostor model. It can be realised that this approach has similarities with UCN. If $\boldsymbol{\lambda}_1^{\mathrm{EI}}, \boldsymbol{\lambda}_2^{\mathrm{EI}}, ..., \boldsymbol{\lambda}_J^{\mathrm{EI}}$ are set to be the cohort speaker models of UCN then the only difference between UCN and the above form of T-norm is the use of the standard deviation.

In the second form, for each operating condition a different set of example impostor models is created. In the test phase, the operating condition of the incoming utterance is determined and the appropriate example impostor set is invoked [7]. Like the first form of Z-norm, this form of T-norm, which requires the explicit knowledge of the operating condition, is not considered in this work.

3.2.3. Adaptation of Z-norm and T-norm

The direct adaptation of the Z-norm and T-norm for open-set identification would result in the following two formulas (only considering the forms which do not require explicit knowledge of the operating condition):

$$L(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}^{\mathrm{ML}} \mid \mathbf{O}) - \boldsymbol{\mu}_{\mathrm{Z}}(\boldsymbol{\lambda}^{\mathrm{ML}})}{\boldsymbol{\sigma}_{\mathrm{Z}}(\boldsymbol{\lambda}^{\mathrm{ML}})}, \tag{14}$$

$$L(\mathbf{O}) = \frac{\log p(\boldsymbol{\lambda}^{\mathrm{ML}} \mid \mathbf{O}) - \boldsymbol{\mu}_{\mathrm{T}}(\mathbf{O})}{\boldsymbol{\sigma}_{\mathrm{T}}(\mathbf{O})}, \tag{15}$$

where all the symbols have the same meanings as before except $\boldsymbol{\mu}_{\mathrm{T}}(\mathbf{O})$ and $\boldsymbol{\sigma}_{\mathrm{T}}(\mathbf{O})$ which are the mean and standard deviation of $\{\log p(\boldsymbol{\lambda}_1|\mathbf{O}), \log p(\boldsymbol{\lambda}_2|\mathbf{O}), \ldots, \log p(\boldsymbol{\lambda}_N|\mathbf{O})\}$.

It should be noted that none of the above adapted versions could lead to a standard form for neither of the two score distributions (i.e. known speaker and unknown speaker score distributions). The reason is that they aim to standardise the distribution of the scores, $\angle(k)$ $k = 1, 2, \ldots$, that would result from $\log p(\boldsymbol{\lambda}_i \mid \mathbf{O}_k)$, $\mathbf{O}_k \notin \boldsymbol{\lambda}_i$ (in the case of Z-norm) or $p(\boldsymbol{\lambda}_k \mid \mathbf{O}_i)$, $\mathbf{O}_i \in \boldsymbol{\lambda}_i$ $(\neq \boldsymbol{\lambda}_k)$, (in the case of T-norm) - in the context of SV, this distribution is the estimated impostor distribution associated the model $\boldsymbol{\lambda}_i$. Figure 2 shows an example of this distribution. However, it should be said that due to the nature of the open-set speaker identification task, in practice, it is considerably difficult to devise a reliable method to transform each form of unknown (or known) speaker score distribution, resulting from a different operating condition, to a standard form. Therefore, for the purpose of this study, it was decided to consider equations (14) and (15) without any modifications. This was also encouraged by the fact that the unknown speaker scores would be part of the distribution of the scores $\angle(k)$, $k = 1, 2, \ldots$.

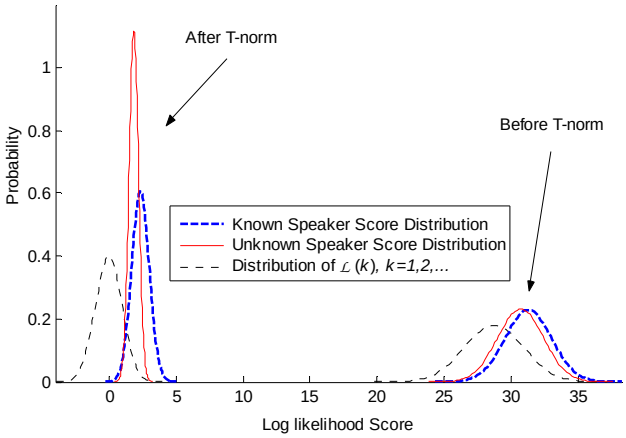

Figure 2: Typical plots of relevant distributions before and after applying T-norm.

4. Experimental investigation

4.1. Speech Data

The speech data adopted for this experimental study was the dataset used for the 1-speaker detection task of NIST Speaker Recognition Evaluation 2003. This dataset consisted of 356 target speakers (207 female and 149 male speakers) and 60 non-target speakers (43 female and 17 male speakers). The training data for each target speaker consisted of two minutes of speech. Each test token consisted of up to 60 seconds of speech from a single speaker.

In order to form a dataset for open set identification, first the target and non-target speakers of the adopted dataset were pooled. This pool was then divided into four subsets. The first subset consisted of 142 speakers (80 female and 62 male speakers) who were enrolled into the system. The second subset was dedicated to the unknown speakers and consisted of 141 (93 females and 48 males) speakers. The third subset contained 100 speakers (58 females and 42 males). This subset was used for the generation of the world model. The forth subset, which was based on 33 speakers (21 females and 12 males) was reserved as a development set and used for the purpose of estimating the Z-norm parameters.

The training data for the known speaker models consisted of two minutes of speech and each test token contained up to 60 seconds of speech from a single speaker. For the purpose of generating the world model, all the training and test utterances were used resulting in about 8 hours of speech material. In the development set a total of 505 utterances were available. In this arrangement there were in total 1293 test utterances for known speakers (767 from females and 526 from males) and 1408 utterances for unknown speakers (893 from females and 515 from males). These test utterances could generate up to 1293 known speaker scores and 1408 known speaker scores for the second stage of the open-set identification. It was felt that such amounts of test scores were inadequate to reliably estimate known and unknown speaker distributions for the purpose of determining OSI-FA and OSI-FR.

In order to obtain more reliable experimental results, it was decided to create another dataset with a composition similar to the one described above, through a data rotation approach. For the purpose of this paper, the above extra dataset is referred to as DataSet 2 whereas the dataset described earlier is referred to as DataSet 1. The first subset of DataSet 2 was formed using all 100 speakers in the third subset of DataSet 1 and 42 randomly selected speakers from second subset of DataSet 1. The remaining 99 (=141 – 42) speakers in the second subset of DataSet 1 were moved to the second subset of DataSet 2. Out of the 142 speakers in first subset of DataSet 1, 100 were randomly chosen to form the third subset of DataSet 2. The remaining 42 speakers were moved to the second subset of DataSet 2. The forth subset of the DataSet 2 is identical to the corresponding one in DataSet 1. Figure 3 illustrates the said data rotation approach. It should be noted that this approach raised the number of known scores to 2563 whilst increasing the maximum number of unknown scores to 2852. Table 1 provides further information on the above sets. It can be easily seen from this table that both datasets have a similar female/male ratio in each of the corresponding subsets.

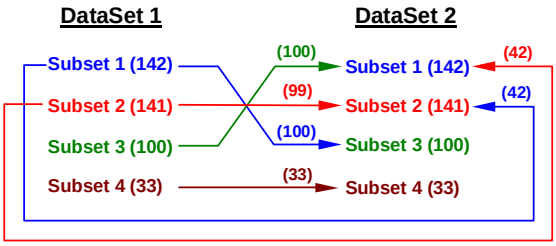

Figure 3: Illustration of the data rotation process.

Speaker sets	*Data Set 1*		*Data Set 2*	
	Female	*Male*	*Female*	*Male*
Known	80	62	80	62
# tests	767	526	728	542
Unknown	93	48	93	48
# tests	893	515	908	536
WM	58	42	58	42
data length	4.8h	3.3h	4.9h	3h
Development	21	12	21	12
# utterances	323	182	323	182

Table 1: Structure of each of the two datasets used in the study.

4.2. Feature Representation

For the purpose of this study, the t^{th} frame of the input speech data was represented as $\mathbf{c}_t \equiv \{c_t(1), c_t(2),\ldots, c_t(16), \Delta c_t(1), \Delta c_t(2),\ldots, \Delta c_t(16)\}$, where $c(i)$ was the i^{th}, mean-subtracted, linear predictive coding-derived cepstral (LPCC) parameter

and $\Delta c(i)$ was the i^{th} delta LPCC parameter. Prior to the extraction of the feature parameters, a simple energy based speech activity detector was used to discard low energy frames. This speech detector discarded between 20-30% of speech frames. The extraction of LPCC parameters was based on first pre-emphasising the input speech data using a first order digital filter and then segmenting it into 20 ms frames at intervals of 10 ms using a Hamming window. $\Delta c_n(i)$ was generated by fitting a linear regression line to $c_{t-2}(i)$, $c_{t-1}(i)$,…, $c_{t+2}(i)$.

4.3. Speaker Representation

In all the experimental investigations discussed in this section, the speaker representation was based on Gaussian mixture models (GMM) [12]. The GMM topologies used to represent each enrolled speaker model and the world model were 32m and 2048m respectively, where *N*m implies *N* Gaussian mixture densities parameterised with a mean vector and diagonal covariance matrices. The parameters of each GMM involved were estimated by using a form of the expectation-maximisation (EM) algorithm [12]. An initial estimate of the model parameters for the EM algorithm was obtained by using a modified version of the LBG procedure termed distortion driven cluster splitting (DDCS) [13]. In order to generate the required world model the following 3 step procedure was followed.

First, two gender dependent models were constructed (using the same EM procedure as that described for the generation of the known speaker models) using clean speech drawn from a down-sampled (16kHz to 8kHz) version of the TIMIT database. All the speech utterances from the 630 speakers (192 females and 438 males) were used for this purpose which accounted for about 4 hours of speech material.

In the second step, a form of Bayesian adaptation [6] was used to adapt each gender dependent world model to the speech data used for this experimental work. For the purpose of this adaptation, the world model subset described in section 4.1 was used.

In the final step a single gender independent world model was created by pooling the adapted model parameters of the two gender dependent 1024m world models. The reason for training gender dependent GMMs was to avoid a gender-based bias developing in the final world model.

4.4. Testing Procedure

For each test trial, first, the following equations were evaluated.

$$S_{\mathrm{ML}} = \max_{1\le n\le N}\left\{\sum_{t=1}^{T}\log p(\mathbf{c}_t \mid \boldsymbol{\lambda}_n)\right\}, \tag{16}$$

$$n_{\mathrm{ML}} = \arg\max_{1\le n\le N}\left\{\sum_{t=1}^{T}\log p(\mathbf{c}_t \mid \boldsymbol{\lambda}_n)\right\}, \tag{17}$$

where $\mathbf{C} \equiv \{\mathbf{c}_1, \mathbf{c}_2,\ldots,\mathbf{c}_T\}$ was the vector sequence representing the test segment, $\boldsymbol{\lambda}_n$ was the GMM representing the n^{th} registered speaker and N was the total number of speakers known to the system. If $\mathbf{C}$ was originated from the m^{th} registered speaker and $n_{\mathrm{ML}} \neq m$ then an OSIE was registered. Otherwise, S_{ML} was normalised (when a score normalisation technique was considered) and stored in one of two groups depending on whether $\mathbf{C}$ 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.

4.5. Experimental conditions, results and discussions

In the case of CN and UCN, experiments were repeated for cohort sizes 1 to 141 (which is the number of registered speakers excluding the $n_{\mathrm{ML}}{}^{th}$). In the CN method, the selection of the competing models was carried out prior to the experiments using a pair-wise comparison technique [4]. In the case of T-norm, the test scores obtained for all the registered speakers were used to compute the relevant mean and standard deviation.

The open-set identification error rate (OSIE) that resulted from this experimental work was about 34%. The OSI-EER's for all the score normalisation methods considered are presented in Figure 4, as a function of the cohort size. The OSI-EER obtained without normalising scores is also given in this figure as the baseline.

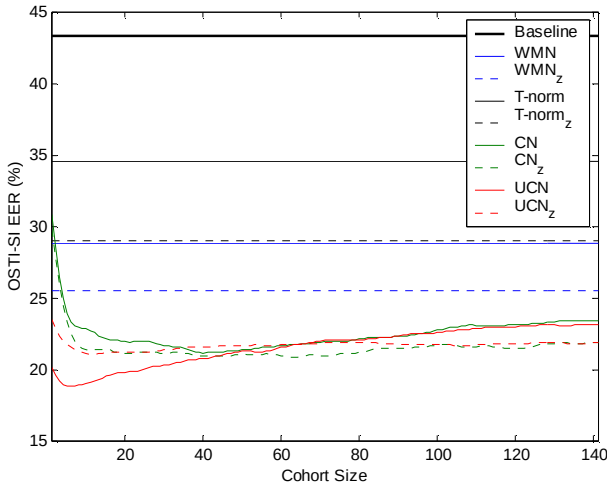

Figure 4: Comparison of various normalisation methods in terms of OSI-EER.

These results clearly confirm the importance of score normalisation in open-set identification. As shown in Figure 4, the lowest OSI-EER in this study was about 18% which was achieved by using UCN. In order to reveal the extent of difficulties in open-set identification, the above experiments were repeated by replacing the open-set identification process with speaker verification. The results showed that the error rates were considerably lower for the latter. For instance, the equal error rate with UCN dropped from 18% in the case of OSTI-SI to only 8% for SV.

Referring to Figure 4 again, it is noted that T-norm is the least effective normalisation method in OSI. This ineffectiveness must be attributed to the way in which T-norm scales the unknown speaker distribution (Section 3.2).

These results also show that, the performance of the WMN method could not exceed that of the cohort approaches except in the case of CN with a cohort of one speaker. In the case of cohort methods, it is noted that the levels of EER obtained were dependent on the cohort sizes used. This suggests that the performance of each of CN and UCN can be optimised by an appropriate selection of their cohort sizes.

It is also interesting to note that as the cohort size is increased, the effectiveness of the CN method improves almost exponentially and the gap between this and the performance of the UCN method decreases. For larger cohort sizes, OSI-EER's obtained using these two methods are almost identical. This is because the cohort sets become increasingly similar.

In general, it can be said that Z-norm produces improvement when combined with all the normalisation methods. The exception to this is the case of Z-norm combined with UCN with small cohort sizes (<50). 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 development set which includes enough varieties of unknown speaker utterances that yield a maximum likelihood for each registered speaker over all registered speakers. Achieving this in practice is almost impossible and therefore, it may be best to avoid the use of combined Z-norm and UCN with small cohort sizes.

In order to analyse the experimental results further, the detection error trade-off curves (DET) for the considered normalisation is presented in Figure 5. It should be noted that for CN and UCN, only the results obtained for the cohort sizes which yield the best performance are shown. The plots clearly indicate the superior performance of cohort methods and, especially, UCN in open-set identification. Comparing UCN with CN, it is observed that, for any operating point fixed according to a given OSI-FA, UCN is more effective in reducing OSI-FR. The best EER for each score normalisation method is also shown in Table 2 to provide further clarification.

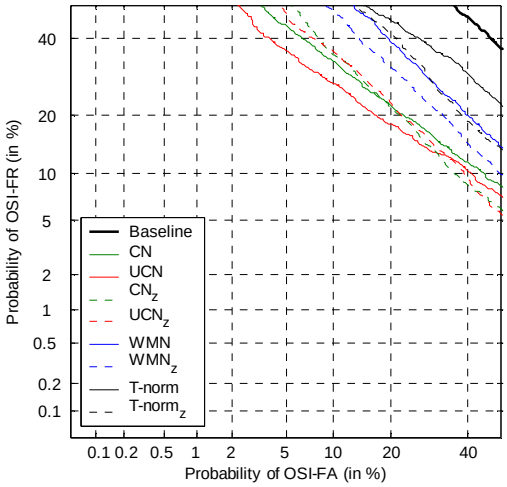

Figure 5: DET curves for various normalisation methods used in OSTI-SI (the cohort sizes chosen for CN and UCN are those that resulted in the best performance).

Normalisation method	Best EER (%)
Baseline	43
T-norm	34
T-norm$_z$	29
WMN	29
WMN$_z$	25
CN	21
UCN$_z$	21
CN$_z$	21
UCN	19

Table 2: Results obtained for the individual normalisation methods considered.

5. Conclusions

The need for score normalisation in the second stage of OSTI-SI has been discussed and two categories of methods that can be used for this purpose have been detailed. The first category, which is based on the Bayesian solution, includes the CN, UCN and WMN methods. The methods in the second category aim to standardise one of the two score distributions involved, and include T-norm and Z-norm. Based on the results of this study it has been observed that the EER in the second stage of OSTI-SI reduces significantly when a score normalisation method is adopted. However the results also indicate that the level of this reduction is dependent on the type of the normalisation method used. It has been shown that for small cohort sizes, CN is outperformed by UCN. As the cohort size is increased, the effectiveness of the CN method improves almost exponentially and, for adequately large

cohort sizes, becomes comparable with that of the UCN method. The experimental results have shown that the WMN method is the least effective method in the first category. It has been demonstrated that the use of Z-norm with WMN and CN improves the performance of these normalisation methods. However, the use of Z-norm with UCN can only be beneficial when the cohort size is adequately large. T-norm has been found to be the overall worst performer for OSI-EER. Although the performance of this method improves considerably when combined with Z-norm, this is still poor compared with methods in the first category.

6. References

[1] Higgins, A., Bahler, L. and Porter, J. "Speaker verification using randomised phrase prompting," *Digital Signal Processing*, vol. 1, pp. 89-106, 1991.

[2] Rosenberg, A. E., DeLong, J., Lee, C. H., Juang, B. H. and Soong, F., "The use of cohort normalised scores for speaker verification," *in Proceedings of the ICSLP'92*, pp. 599-602, 1992.

[3] Reynolds, D. A., "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, August 1995.

[4] Rosenberg, A. E. and Parthasarathy, S. "Speaker background models for connected digit password speaker verification," *in Proceedings of the ICASSP'96*, pp. 81-84, 1996.

[5] Ariyaeeinia, A. M. and Sivakumaran, P., "Analysis and comparison of score normalisation methods for text-dependent speaker verification," *in Proceedings of the Eurospeech'97*, pp. 1379-1382.

[6] Reynolds, D. A., "Comparison of background normalisation methods for text-independent speaker verification," *in Proceedings of the Eurospeech'97*, pp. 963-966.

[7] 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.

[8] Gish, H. and Schmidt, M., "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18-32, October 1994.

[9] Markov, K. and Nakagawa, S., "Text-independent speaker recognition using non-linear frame likelihood transformations," *Speech Communication*, vol 24, pp. 193-209, 1998.

[10] Rosenberg, A., Parthasarathy, S., Hirschberg, J. and Whittaker, S., "Foldering voicemail messages by caller using text independent speaker recognition," *in Proceedings of the ICSLP'00*, 2000.

[11] Viswanathan, M., Beigi, H. S. M, Dharanipragada, S., Maali, F. and Tritchler, A., "Multimedia document retrieval using speech and speaker recognition," *International Journal on Document Analysis and Recognition*, 2(4): pp. 147-162, 2000.

[12] Reynolds, D. A., "Robust text-independent speaker identification using gaussian mixture models," *IEEE Transactions on Speech and Audio Processing*, pp. 72-83, vol. 3, no. 1, January 1995.

[13] 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.