

Score Normalisation Applied to Open-Set, Text-Independent Speaker Identification

P. Sivakumaran, J. Fortuna* and A. M. Ariyaeinia*

20/20 Speech Ltd., Malvern, Worcestershire, WR14 3SZ, UK

*University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK

p.sivakumaran@2020speech.com, {j.m.r.c.fortuna, a.m.ariyaeinia}@herts.ac.uk

Abstract

This paper presents an investigation into the relative effectiveness of various score normalisation methods for open-set, text-independent speaker identification. The paper describes the need for score normalisation in this case, and provides a detailed theoretical and experimental analysis of the methods that can be used for this purpose. The experimental investigations are based on the use of speech material drawn from 9 hours of recordings of different Broadcast News. The results clearly demonstrate the significance of improvement offered by score normalisation. It is shown that, amongst various normalisation methods considered, the unconstrained cohort normalisation method achieves the best performance in terms of reducing the errors associated with the open-set nature of the process. Furthermore, it is demonstrated that both the cohort and world model methods can offer very similar effectiveness, and also outperform the T-norm method in this particular case of speaker recognition.

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 exasperate 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, and has a range of potential applications. Examples are surveillance, forensics and document retrieval.

The inherent complexity of OSTI-SI is primarily attributed to the stage where the decision is made to declare the test utterance as not belonging to any of the known speakers. The reason is that this stage is highly susceptible to 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 which in turn lead to a significant reduction in the effectiveness of the system. 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 [1-7]. However, the nature of the problem here is somewhat different from that of speaker verification and therefore, it is

not possible to foresee the outcome of this study from those of the speaker verification studies.

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 database used for the experimental investigation and provides details on the adopted speaker modelling methods. The experimental work together with the results 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 $\lambda_1, \lambda_2, \dots, \lambda_N$. If \mathbf{O} denotes the feature vectors sequence extracted from the test utterance then the open-set identification can be stated as follows:

$$\max_{1 \leq n \leq N} \{p(\mathbf{O} | \lambda_n)\} \geq \theta \rightarrow \mathbf{O} \in \begin{cases} \arg \max_{1 \leq n \leq N} \{p(\mathbf{O} | \lambda_n)\} \\ \text{unknown speaker model} \end{cases} \quad (1)$$

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 the speaker models in the system, if the 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 errors are possible:

- \mathbf{O} , which belongs to λ_m , not yielding the maximum likelihood for λ_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, and yield the maximum likelihood for, λ_m as originated from an unknown speaker.

For the purpose of this paper these 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).

Further to the discussions in Section 1, and based on equation (1), the two-stage process in open-set identification can be reiterated as follows. 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.

The important point to note in this two-stage process is that the latter stage is far more susceptible than the former stage to distortions in test speech characteristics. 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 reflected in all the likelihood scores. As a conse-

quence, 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. In particular, score normalisation techniques are considered here to tackle this problem. The details of these methods and related issues are discussed in the next section.

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 (SV). However, in the case of SV, 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 corresponding score distributions in speaker verification.

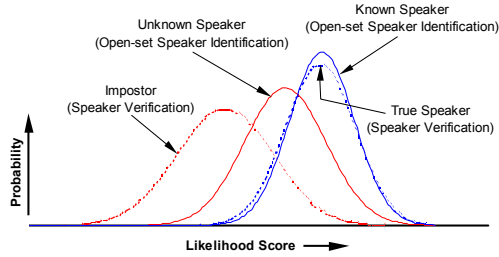


Figure 1: Score distributions associated with the speaker verification and the second stage of open-set speaker identification.

3. Score normalisation

3.1. Bayesian solution

In the probabilistic view, the decision rule for the second stage of open-set identification can be expressed as follows.

$$P(\lambda^{\text{ML}} | \mathbf{O}) \geq P(\lambda^{\text{U}} | \mathbf{O}) \rightarrow \mathbf{O} \in \begin{cases} \lambda^{\text{ML}} \\ \lambda^{\text{U}} \end{cases} \quad (2)$$

where $\lambda^{\text{ML}} = \arg \max_{1 \leq n \leq N} \{p(\mathbf{O} | \lambda_n)\}$ 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

$$\frac{p(\mathbf{O} | \lambda^{\text{ML}})}{p(\mathbf{O} | \lambda^{\text{U}})} \geq \frac{P(\lambda^{\text{U}})}{P(\lambda^{\text{ML}})} \rightarrow \mathbf{O} \in \begin{cases} \lambda^{\text{ML}} \\ \lambda^{\text{U}} \end{cases} \quad (3)$$

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

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

In order to realise the benefit of this Bayesian solution fully, $p(\mathbf{O} | \lambda^{\text{U}})$ has to be determined accurately. However, λ^{U} which represents the model for unknown speakers is unavailable in practice. Therefore, the best option is to determine an appropriate replacement for $p(\mathbf{O} | \lambda^{\text{U}})$ 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 for making the final decision [2,5]:

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

where λ^{C} is the model associated with the claimed identity and, λ^{I} is the impostor model which is, in fact, unavailable in practice. In speaker verification, in order to tackle this problem, various techniques have already been proposed [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.

- **World model normalisation (WMN)**

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

- **Cohort normalisation (CN)**

In this method, each enrolled speaker is associated with a cohort of speakers whose models are most competitive with the model of that enrolled speaker. 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 $\log p(\mathbf{O} | \lambda^{\text{U}})$ is computed as:

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

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

- **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, $\log p(\mathbf{O} | \lambda^{\text{U}})$ is replaced with

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

where, $\phi(i) \neq \phi(j)$ if $i \neq j$ and, $\lambda_{\phi(1)}, \lambda_{\phi(2)}, \dots, \lambda_{\phi(K)}$ are the models which yield the next K highest likelihood scores to $p(\mathbf{O} | \lambda^{\text{ML}})$. This method can also be viewed as a special case of CN method where the required cohort of speakers is chosen according to their closeness to the test utterance.

The overall effectiveness of the above score normalisation methods depends on two criteria: (1) ability to compensate for distortions in the test utterance, (2) ability to produce the unknown speaker scores that are smaller than the known speaker scores. The CN method is the primary candidate to deal with the first criterion. This is because, it is expected that the cohort models which are highly competitive with λ^{ML} , would better replicate the way in which $p(\mathbf{O} | \lambda^{\text{ML}})$ is affected by the distortions in the test utterance. It is also expected that the UCN method to be almost as effective as the CN method in dealing with first criterion. This is due to the fact that, if $\mathbf{O} \in \lambda_m$ and $\lambda^{\text{ML}} = \lambda_m$, the cohort speaker models chosen in the UCN method using \mathbf{O} , would be almost the same as the cohort speaker models chosen in the CN method based on their closeness to λ_m . This implies that the relative effectiveness of the CN and UCN methods is largely dependent on second criterion.

The WMN method is expected to be very competitive with the UCN method. The reason is that the world model encodes each group of similar speakers into the same set of mixtures (it is assumed that the world model, λ^{WM} , is represented by using Gaussian mixtures). As a result, in the generation of $p(\mathbf{O} | \lambda^{\text{WM}})$ there are heavier contributions by sets of mixtures that encoded the groups of speakers to whom \mathbf{O} is closer.

3.2. Standardising a score distribution

The methods in this category, like the score normalisation methods described in the previous section, are originally proposed for speaker verification. In their original form, they aim to transform each form of the impostor score distribution, resulting from a different test condition, to a standard form. 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 transformation parameters. Further details of the two main methods in this category are given below. In both of these methods, the impostor score distributions are assumed to be Gaussian.

• Zero normalisation (Z-norm)

In this case, the score normalisation is performed according to the following equation

$$L_{sv}(\mathbf{O}) = \left\{ \log p(\mathbf{O} | \lambda^c) - \mu_z(\lambda^c, \varphi(\mathbf{O})) \right\} / \sigma_z(\lambda^c, \varphi(\mathbf{O})) \quad (8)$$

where $\mu_z(\cdot, \cdot)$ and $\sigma_z(\cdot, \cdot)$ are specific to the model associated with the claimed identity and represent the mean and standard deviation of the impostor score distribution for the operating condition given by $\varphi(\mathbf{O})$. In the training phase, the pair $\{\mu_z(\cdot, \cdot), \sigma_z(\cdot, \cdot)\}$ is computed for each registered speaker in each considered operating conditions using a set of development impostor utterances. In this method, a scheme has to be devised to detect the operating condition in the test phase [6]. It should be noted that this normalisation technique has been successfully applied to tackle the problem caused by handset mismatch in speaker verification (in this case, the method is specifically known as handset normalisation or *H-norm* [6]). One obvious weakness of this method is that it cannot be used effectively in an unknown operating condition.

• Test normalisation (T-norm)

This method does not require any explicit knowledge of the operating condition. It uses a set of example impostor models to determine the required parameters in the test phase, more formally:

$$L_{sv}(\mathbf{O}) = \left\{ \log p(\mathbf{O} | \lambda^c) - \mu_t(\mathbf{O}) \right\} / \sigma_t(\mathbf{O}) \quad (9)$$

where $\mu_t(\mathbf{O})$ and $\sigma_t(\mathbf{O})$ are the mean and standard deviation of $\log p(\mathbf{O} | \lambda_1^{\text{EGI}})$, $\log p(\mathbf{O} | \lambda_2^{\text{EGI}})$, ..., $\log p(\mathbf{O} | \lambda_j^{\text{EGI}})$ and, λ_j^{EGI} is the j^{th} example impostor model. It can be realised that this approach has similarities to UCN. The main difference here is the use of the standard deviation.

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

$$L(\mathbf{O}) = \left\{ \log p(\mathbf{O} | \lambda^{\text{ML}}) - \mu_z(\lambda^{\text{ML}}, \varphi(\mathbf{O})) \right\} / \sigma_z(\lambda^{\text{ML}}, \varphi(\mathbf{O})) \quad (10)$$

$$L(\mathbf{O}) = \left\{ \log p(\mathbf{O} | \lambda^{\text{ML}}) - \mu_t(\mathbf{O}) \right\} / \sigma_t(\mathbf{O}) \quad (11)$$

where all the symbols have the same meanings as before except $\mu_t(\mathbf{O})$ and $\sigma_t(\mathbf{O})$ which are the mean and standard deviation of $\{\log p(\mathbf{O} | \lambda_1), \log p(\mathbf{O} | \lambda_2), \dots, \log p(\mathbf{O} | \lambda_N)\}$.

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 scores, $\mathcal{L}(k)$ $k=1,2,\dots$, that would result from $\log p(\mathbf{O} | \lambda_i)$ $\mathbf{O} \neq \lambda_j$ (Figure 2). However, it should be said that due to the nature of the open set speaker identification task, in practice, it is impossible to devise a reliable method to transform each form of unknown (or known) speaker score distribution, resulting from a different test condition, to a standard form. Therefore, for the purpose of this study, it was decided to consider equations (10) and (11) without any modifications. This was also motivated by the fact that the unknown speaker scores would be part of the distribution of the scores $\mathcal{L}(k)$, $k=1,2,\dots$

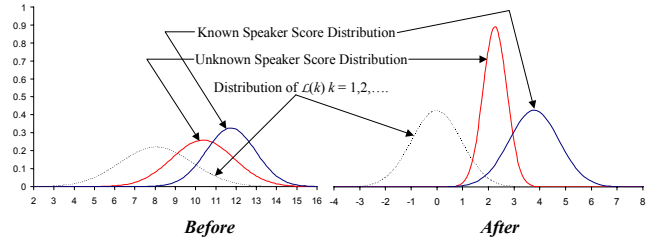


Figure 2: Typical plots of relevant distributions before and after applying Z/T-norm.

4. Experimental investigation

4.1. Speech data

The speech data used in the experimental study was drawn from 9 hours of recordings of different Broadcast News. It consisted of two subsets, both sampled at a rate of 16 kHz. The first subset was used for the purpose of training a world model. This subset consisted of 4 hours of speech data corresponding to 500 (275 male & 225 female) speakers. The second subset consisted of 125 (65 male & 60 female) speakers in which 68 (36 male & 32 female) speakers were enrolled into the system and the remaining 57 were set to act as unknown speakers. For each enrolled speaker, 3 minutes of speech data was available, in which 1.5 minutes of speech data was used for building the speaker models and the remaining 1.5 minutes was reserved for testing. The length of the test speech data reserved for each unknown speaker was also 1.5 minutes. It should be noted that care was taken to ensure that the speech data used in training and testing were drawn from different recordings.

4.2. Feature parameter representation

For the purpose of this study, the i^{th} frame of the input speech data was represented as $\mathbf{c}_i \equiv \{c_i(1), c_i(2), \dots, c_i(20), \Delta c_i(1), \Delta c_i(2), \dots, \Delta c_i(20)\}$, where $c_i(i)$ is the i^{th} , mean subtracted, linear predictive coding-derived cepstral (LPCC) parameter and $\Delta c_i(i)$ is the i^{th} delta LPCC parameter. The extraction of LPCC parameters was based on first pre-emphasising the input speech data using a first order digital filter and then segmenting into 30 ms frames at intervals of 15 ms using a Hamming window. $\Delta c_n(i)$ was generated by fitting a linear regression line to $c_{i-2}(i), c_{i-1}(i), \dots, c_{i+2}(i)$.

4.3. Speaker representation

In all the experimental investigations discussed in this section, the speaker representation was based on the Gaussian mixture models (GMM). The GMM topologies used to represent each enrolled speaker and the world model were 32m and 2048m respectively, where Nm implies N Gaussian mixture densities which are parameterised with diagonal covariance matrices. The parameters of each GMM involved were estimated by using a form of the expectation-maximisation (EM) algorithm [3].

4.4. Testing procedure

For each test trial, the following equations are evaluated first:

$$S_{ML} = \max_{1 \leq n \leq N} \left\{ \sum_{t=1}^T \log p(\mathbf{c}_t | \gamma_n) \right\} \quad (12)$$

$$n_{ML} = \arg \max_{1 \leq n \leq N} \left\{ \sum_{t=1}^T \log p(\mathbf{c}_t | \gamma_n) \right\} \quad (13)$$

where $\mathbf{C} \equiv \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_T\}$ is the vector sequence representing the test segment, γ_n is the GMM representing the n^{th} registered speaker and N is the total number of speakers known to the system. If C is originated from the m^{th} registered speaker and $n_{ML} \neq m$ then it is assumed that an OSIE has occurred. Otherwise, S_{ML} is normalised (if a score normalisation technique is considered) and is stored in one of two groups depending on whether C is originated from a known or an unknown speaker.

After the completion of all the test trials in a given investigation, the stored S_{ML} values are retrieved to form the empirical score distributions of the known and unknown speakers. These distributions are then used to determine the open-set identification equal error rate (OSI-EER) i.e., the probability of equal number of OSI-FA and OSI-FR.

4.5. Experimental conditions, results and discussions

For the purpose of the experimental investigation, it was decided to consider only the score normalisation methods that tackle the effect of mismatch between the training and test condition without any explicit knowledge of the nature of the mismatch (in other words, all the score normalisation techniques described in Section 3 except the Z-norm). In the case of CN and UCN methods, experiments were repeated for cohort sizes 1 to 67 (which is the number of registered speakers excluding n_{ML}). In the CN method, the selection of the competing models was carried out using a pair-wise comparison technique [2]. In the case of T-norm, the test scores obtained for all the registered speakers were used to compute the relevant mean and the standard deviation. Furthermore, the utilised world model was formed by using two independently generated 1024m GMMs, each representing a gender. In this investigation, an OSIE rate of 3.9% was obtained irrespective of the score normalisation method considered. Figure 3 shows the OSI-EER for all the considered score normalisation methods as a function of cohort size. The OSI-EER obtained using unnormalised scores is also presented in this figure as the baseline.

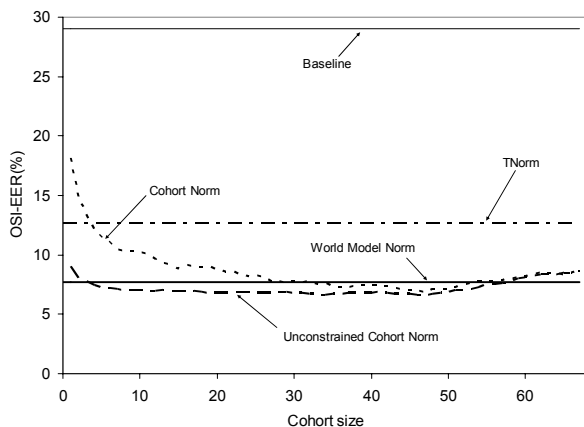


Figure 3: Comparison of various methods in terms of OSI-EER.

It is clear from these results how important is the score normalisation for open-set identification. The lowest OSI-EER of 6.6% is obtained for the UCN method. However, it is evident from these results that an OSI-EER value very close to this can be achieved by all the other considered score normalisation methods except the T-norm. The relative ineffectiveness of the T-norm must be attributed to the way in which it scales the unknown speaker distribution (Section 3.2). 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, the OSI-EERs obtained using these two methods are almost identical. The poor performance of the CN method for smaller cohort sizes must be due its inability in reducing the normalised scores of the unknown speakers.

In order to analyse the experimental results further, the detection error trade-off curves (DET) for all considered methods are given in Figure 4 (in the case of CN/UCN, only the results obtained for the cohort size which yield the typical best performance are shown). It is observed that the CN/UCN and WMN

methods which achieve very close OSI-EERs, give significantly different OSI-FR rate in the regions of lower/higher OSI-FA rates. This implies that the level of FA/FR is important in choosing normalisation method.

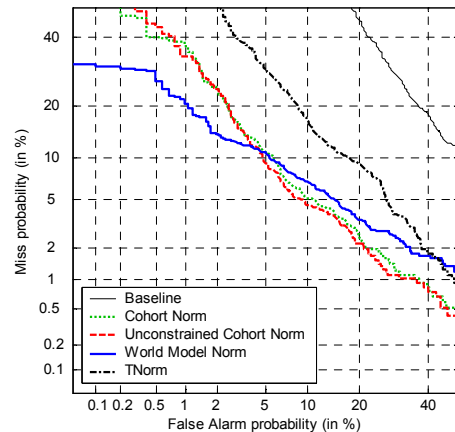


Figure 4: DET curves for various methods (the cohort sizes chosen for CN and UCN are 45 and 5 respectively).

5. Conclusions

The need for score normalisation in OSI was discussed and two groups of methods that can be used for this purpose were detailed. The first group, which was based on the Bayesian solution, included the CN, UCN and WMN methods. The second group, which was aimed to standardise one of the two score distributions involved, included the T-norm and Z-norm. It was shown that the effectiveness of the CN method could not exceed that of the UCN method. In fact, it was observed that for smaller cohort sizes the CN method performed significantly worse than the UCN method. When the cohort size was increased, the effectiveness of the CN method improved almost exponentially and began to converge with that of the UCN method. As expected theoretically, the WNM method achieved a performance very close to that of UCN. The difficulty in using the T/Z-norm method for OSI was also discussed and experimentally demonstrated. Moreover, it was shown experimentally that the level of OSI-FA/FR is important in choosing the normalisation method for the purpose of OSI. All the experiments carried out were based on the use speech material drawn from 9 hours of recordings of different Broadcast News.

6. References

- [1] Higgins, A., *et al*, "Speaker verification using randomised phrase prompting," *Digital Signal Processing*, vol. 1, pp. 89-106, 1991.
- [2] Rosenberg, A. E., *et al*, "The use of cohort normalised scores for Speaker Verification," in *Proc. 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 *Proc. 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 *Proc. Eurospeech'97*, pp. 1379-1382.
- [6] Reynolds, D. A., "Comparison of background normalisation methods for text-independent speaker verification," in *Proc. Eurospeech'97*, pp. 963-966.
- [7] Auckenthaler, R., *et al*, H., "Score normalisation for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.