Effectiveness in Open-Set Speaker Identification

Rawande Karadaghi, Heinz Hertlein and Aladdin Ariyaeeinia
University of Hertfordshire
Hatfield, UK
{r.karadaghi, h.hertlein, a.m.ariyaeeinia}@herts.ac.uk

Abstract— **This paper presents investigations into the relative effectiveness of two alternative approaches to open-set text-independent speaker identification (OSTI-SI). The methods considered are the recently introduced i-vector and the more traditional GMM-UBM method supported by score normalisation. The study is motivated by the growing need for effective extraction of intelligence and evidence from audio recordings in the fight against crime. OSTI-SI is known to be the most challenging subclass of speaker recognition, and its adoption in criminal investigation applications is further complicated by undesired variations in speech characteristics due to changing levels of environmental noise. In this study, the experimental investigations are conducted using a protocol developed for the identification task, based on the NIST speaker recognition evaluation corpus of 2008. In order to closely cover relevant conditions in the considered application areas and investigate the identification performance in such scenarios, the speech data is contaminated with a range of real-world noise. The paper provides a detailed description of the experimental study and presents a thorough analysis of the results.**

Keywords—Open-set speaker identification; GMM-UBM, i-vector

I. INTRODUCTION

Open-set speaker identification is the process of determining the correct speaker of a given utterance from a registered population, with the additional requirement to establish if the utterance is not produced by any of the registered speakers. When the speakers are not required to provide utterances of specific texts during identification trials, the process is referred to as open-set, text-independent speaker identification (OSTI-SI). This is the most challenging class of voice biometrics and has a wide range of applications in such areas as audio surveillance, document indexation, and screening [1].

The past several years have witnessed considerable research into enhancing the effectiveness of open-set speaker identification in practical applications. An aspect of this has been related to the introduction of methods for minimising the adverse effects of speech variation due to additive noise [2, 3]. The significance of this is due to the fact that in practice, additive noise causes a mismatch between the test and reference utterances, which in turn can significantly reduce the reliability of OSTI-SI. This can potentially limit the usefulness of the process, especially in applications where there is little or no control over the noise conditions.

The aim of this study is to investigate the effects of environmental noise on the accuracy of open-set speaker recognition. The speaker classification approaches considered for the purpose of experiments are (i) the state-of-the-art i-vector method and (ii) the traditional GMM-UBM method supported by score normalisation. In order to closely cover relevant conditions in the considered application areas and investigate the effects on the identification performance in such scenarios, the speech data is contaminated with different types of real-world noise.

The remainder of this paper is organised as follows. The next section provides an overview of the approaches to speaker identification adopted in this study. Section III presents the experimental investigations together with an analysis of the results. Finally, the overall conclusions and future work are discussed in Section IV.

II. ADOPTED APPROACHES

A. GMM-UBM with score normalisation

As the well-known GMM-UBM technique has been one of the dominating approaches in the field of speaker recognition for the past two decades [4, 5], it is considered as the baseline in the experimental part of this study.

The recognition accuracy of the traditional GMM-UBM approach can be improved significantly by applying additional score normalisation techniques. The improvement in this case is achieved by increasing the separation between the score distributions for known and unknown speakers. The normalisation approaches adopted for the purpose of this study are T-norm and Z-norm, which are based on the standardisation of score distributions [1, 6].

B. The i-vector total variability space

The i-vector approach [3, 7] is related to the GMM-UBM technique as each i-vector can be regarded as a compact representation of an adapted GMM. To this end, a matrix $\boldsymbol{T}$ called total variability matrix is computed from a large background corpus. The name total variability matrix refers to the fact that in i-vector space, speaker specific information is contained together with intra-speaker variability. This matrix $\boldsymbol{T}$ defines a transformation of GMM Gaussian mean supervectors to the lower-dimensional i-vector space:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x} \quad (1)$$

Here, $\boldsymbol{M}$ is the means supervector corresponding to the speech utterance, $\boldsymbol{m}$ is the UBM supervector and $\boldsymbol{x}$ is a standard-normally distributed latent variable of the dimension chosen for the i-vector space. The i-vector $\boldsymbol{w}$ that represents

978-1-4799-3532-1/14/$31.00 ©2014 IEEE

the speech utterance is computed as the MAP estimate of $\boldsymbol{x}$ [8]. As noted above, the total variability matrix $\boldsymbol{T}$ is computed as ML estimate from a background corpus. This corpus should be sufficiently large and representative of the speech conditions encountered in relevant applications.

Unlike traditional GMM-UBM score computation, the i-vector approach is symmetrical in the sense that i-vectors are computed for training and test utterances. The comparison of the test i-vector, $\boldsymbol{w}_{test}$, and target i-vector, $\boldsymbol{w}_{target}$, is conducted by using the cosine similarity score (CSS) defined as follows [3]:

$$\text{score}(\boldsymbol{w}_{target}, \boldsymbol{w}_{test}) = \frac{< \boldsymbol{w}_{target}, \boldsymbol{w}_{test} >}{\|\boldsymbol{w}_{target}\| \, \|\boldsymbol{w}_{test}\|}. \qquad (2)$$

Due to the fact that i-vectors represent not only the characteristics of the speaker that is important for the recognition task, but also undesired intra-speaker variability such as channel effects, the suppression of the latter improves the accuracy of the approach [9]. To that end, the technique of within-class covariance normalisation (WCCN) has been shown to work well in practice. To apply this technique, a covariance matrix is computed for each one of the speakers in a background set. Then, the average of all these covariance matrices is calculated to obtain the overall within-class covariance matrix $\boldsymbol{W}$ [3]:

$$\boldsymbol{W} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\boldsymbol{w}_i^s - \overline{\boldsymbol{w}_s})(\boldsymbol{w}_i^s - \overline{\boldsymbol{w}_s})^t. \qquad (3)$$

Here, $\overline{\boldsymbol{w}_s}$ is the mean of all i-vectors in the set originating from speaker s ($s = 1, \ldots, S$) and $\boldsymbol{w}_i^s$ is the i^{th} i-vector of speaker s in the background set ($i = 1, \ldots, n_s$). Then, a matrix $\boldsymbol{B}$ is obtained through Cholesky decomposition of the inverse of the within-class covariance matrix; $\boldsymbol{W}^{-1} = \boldsymbol{B}\boldsymbol{B}^t$. Finally, matrix $\boldsymbol{B}$ is multiplied with any i-vector $\boldsymbol{w}$ to calculate its normalized version:

$$\boldsymbol{w}_{\text{norm}} = \boldsymbol{B}^t \boldsymbol{w}. \qquad (4)$$

III. Experimental Investigations

A. Speech corpora and the protocol for the evaluation of OSTI-SI

The experiments in this paper are mostly based on the NIST speaker recognition evaluation (SRE) database 2008. An evaluation protocol for open-set identification has been defined on a subset of this telephone-quality database containing 400 registered speakers and 200 out-of-set (unknown) speakers. All the selected material originates from the "short2/short3" core condition [10]. The number of identification trials depends not only on the number of registered speakers and out-of-set impostors, but also on the number of test utterances which varies for different speakers. For this reason, there are a total of 1312 identification trials of enrolled speakers and 627 identification trials of out-of-set impostors.

The background corpus used for UBM training is a subset of the NIST speaker recognition evaluation (SRE) database 2005 [11]. This dataset consists of 622 male and 932 female utterances, and the developed gender independent UBM comprises 2048 Gaussian mixture components.

B. Adopted approaches for speaker classification

A total of three speaker recognition techniques are included in the experimental investigations:

(i) GMM-UBM as baseline system, without additional score normalization techniques
(ii) GMM-UBM as in (i), but with TZ-norm [12, 13]
(iii) The i-vector approach, with a dimension of 300 for the total variability space.

C. The conditions of noise contamination

Real-world applications of OSTI-SI should be able to cope with a variety of noise types and various degrees of severity of speech signal degradation. Therefore, in order to thoroughly investigate the effect of ambient noise on the OSTI-SI accuracy, a total of seven conditions have been considered in the experiments. The first operating condition is based on the use of the original telephone data from the NIST corpus 2008, without any additional noise contamination. Then, the same speech data has been contaminated with white noise, car and factory noise from the NOISEX-92 corpus of noise recordings [14]. For each one of these three noise types, two versions of the speech corpus have been generated with signal-to-noise ratios (SNRs) of 5 dB and 15 dB, respectively. It should be noted that for each set of experiments, noise of the same type and level has been added to training and test material, and the same type and level has also been added to the background corpus for TZ normalization. The background set of speech utterances for UBM and i-vector total variability training, on the other hand, is not contaminated with noise. This is because it is considered unfeasible in practice to adapt this large part of the background corpus to changing conditions of the speech data and repeat the computationally demanding processes for UBM and total variability matrix generation each time a new condition is encountered.

D. Overview of results for the first stage of OSTI-SI

It is worth noting that the open-set, text-independent speaker identification (OSTI-SI) process consists of the two stages of identification and verification. The accuracy of the first stage can be expressed as the identification rate in the closed-set mode. This accuracy rate is essentially computed based on the use of speech data from the 400 registered population. In other words, the unknown speakers (out of set) cannot influence the results for this stage.

Table 1 gives an overview of the identification (closed-set) rate, for the three classification approaches adopted and the seven noise conditions as defined above. As observed in this table, background noise has a severe effect on the recognition accuracy of the first stage of OSTI-SI. Comparing results of different noise levels at the same approach and the same noise type, unsurprisingly, the drop in identification rate at the lower SNR of 5 dB is significantly larger than that at 15 dB. Table 1 also shows that there are considerable differences between the identification rates for different types of noise (same SNR). As

indicated in this table, white noise has the largest effect, followed by factory noise.

TABLE I

		Identification rate		
		GMM-UBM	GMM-UBM TZnorm	I-Vector
Clean data		39.7%	42.5%	49.5%
White noise contamination	5db	14.7%	19.8%	27.1%
	15db	24.6%	29.7%	39.3%
Car noise contamination	5db	32.1%	37.7%	41%
	15db	34.8%	40.3%	44%
Factory noise Contamination	5db	22.3%	26.3%	33.8%
	15db	30%	33.4%	43%

E. The accumulated error rate (AER)

In order to evaluate the full process of OSTI-SI, the recognition accuracy will need to be computed separately for the two stages of identification and verification. Based on this evaluation strategy, the identification rates are as given in Table 1 above. However, this is not the optimal approach if it is required to compare the effectiveness of different OSTI-SI techniques. In this case, it is more convenient to adopt a single measure of accuracy that characterises the whole process of open-set identification. Motivated by the approach proposed for the computation of error rate in the diarisation process [15], such a measure for OSTI-SI has been introduced in [16]. This measure is referred to as accumulated error rate (AER) and it allows the evaluation of both stages of OSTI-SI using a single error figure. This is defined as the ratio of the sum of inaccuracies encountered over the total number of identification trials:

$$AER(\theta) = \frac{ML(\theta) + FR(\theta) + FA(\theta)}{T}. \qquad (5)$$

Here, θ is the threshold adopted in the second stage of the process, $ML(\theta)$ is the number of mislabeled (incorrectly identified) clients, $FR(\theta)$ is the number of clients that are falsely rejected, $FA(\theta)$ is the number of impostors that are falsely accepted as clients, and T is the total number of identification trials.

This measure of OSTI-SI accuracy is used in the remainder of this paper for a more thorough analysis of the experimental results.

F. AER results with telephone-quality data

In this part of the experimental investigations, the original training and testing data from the NIST evaluation 2008 is used without any additional noise contamination as baseline result. For this condition, Figure 1 shows the accumulated error rate (AER), as defined above, versus the threshold θ. It should be noted that the plots given in this figure are based on applying score range normalisation to the AERs for the three considered methods. This is to facilitate a meaningful comparison of the methods. However, it is noted that in each case, a different threshold still needs to be set in order to achieve the minimum AER. The reason for this is that *AER(θ)* depends on the method-specific client and impostor score distributions. For the purpose of facilitating the comparison further, an extended procedure for score range normalisation is applied to the plots in Figure 2 (and in all subsequent figures), in order to shift the point of minimum AER to the same threshold θ=0.5. Being in the middle of the score range, this value has been chosen to facilitate the graphical representation.

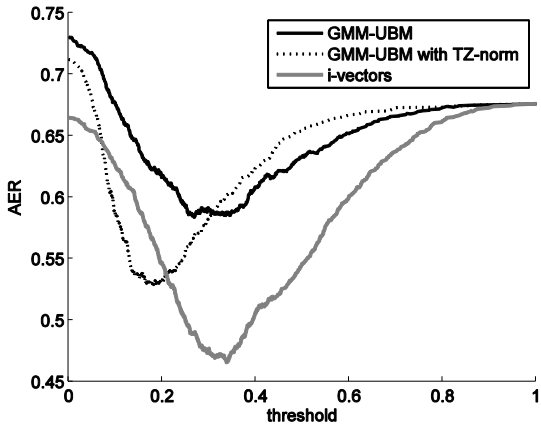

Figure 1: Comparison of different methods based on the NIST telephone-quality speech data.

The experimental results for the NIST telephone-quality data show that the accuracy in OSTI-SI based on GMM-UBM can be considerably improved by using TZ-normalisation. It is also noted that the highest accuracy in this case is offered by the i-vector approach.

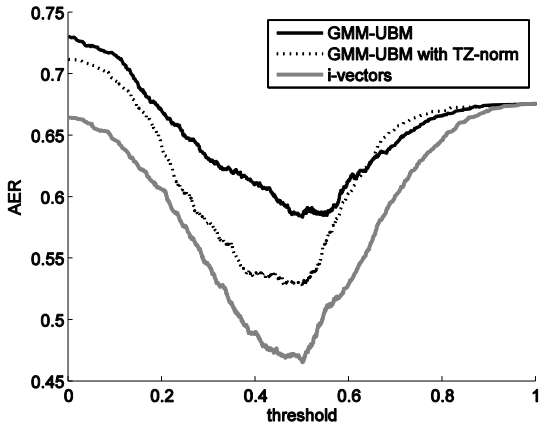

Figure 2: Adjusted AER plots for the experiments in the first part of the investigations.

G. AER results for speech data contaminated with white noise

The aim of the experiments in this and the following section is to comparatively evaluate the recognition performance of the adopted algorithms for different types and levels of noise in speech signals. The speech data contamination in this part is based on the procedure described in section III.C.

The first part of the investigations in this section is based on using white noise to contaminate speech to achieve signal-to-noise ratios (SNRs) of 5 dB (Figure 3) and 15 dB (Figure 4).

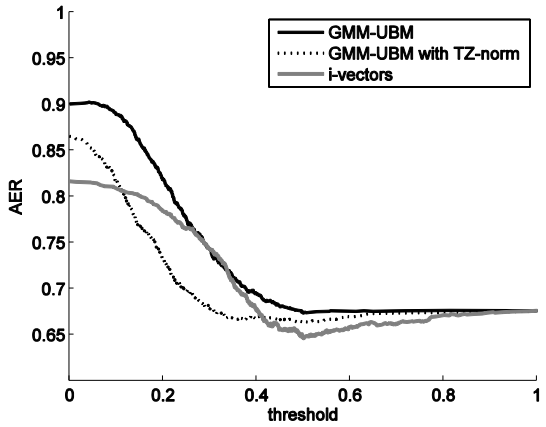

Figure 3: Experimental results for different methods based on the use of speech data contaminated with a high level of white noise (SNR = 5 dB).

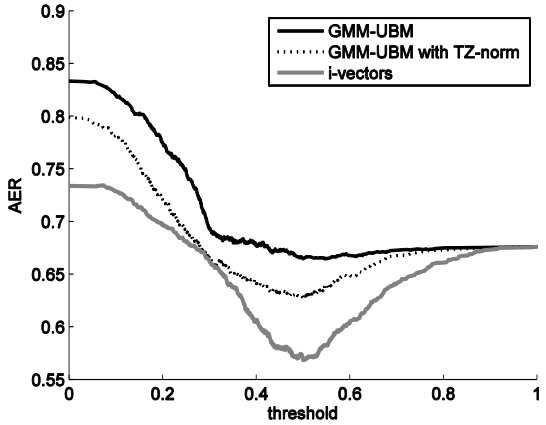

Figure 4: AER plots for different methods based on the use of speech data contaminated with a moderate level of white noise (SNR = 15 dB).

The plots in figures 3 and 4 show that in the case of the lower SNR (5 dB), there is little difference between the three considered approaches as far as the minimal AER is concerned. This is in spite of the improvement of the identification rate in the first stage that is achieved by TZ-normalisation and i-vectors respectively in comparison with the GMM-UBM baseline. However, when the SNR is increased to 15 dB, the i-vector and GMM-UBM with TZ-norm offer higher accuracy rates than the baseline system (Figure 4).

H. AER results for speech data contaminated with car and factory noise

In order to more realistically reflect the conditions encountered in real applications, the experimental investigations are extended to include car and factory noise. As in the previous section, SNRs of 5 and 15 dB are considered for both types of noise. The experimental results for the resultant four conditions are presented in figures 5 to 8.

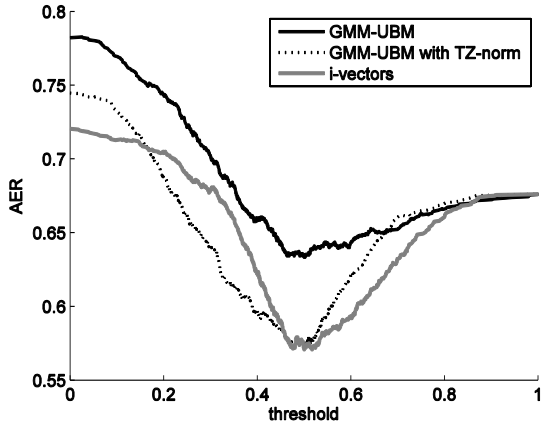

Figure 5: AER plots for speech data contaminated with car noise (SNR = 5 dB).

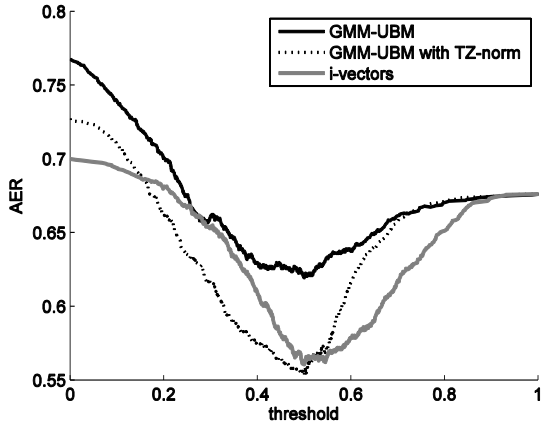

Figure 6: Experimental results for speech data that is moderately contaminated with car noise (SNR = 15 dB).

There are a number of interesting observations to be made from these results. For instance, it can be seen that the synthetic white noise has a more severe adverse effect on OSTI-SI accuracy in comparison with the real-world noise types. Moreover, when comparing the minimal AERs for the different classification techniques considered, it can be noted that in the case of car noise, there is little difference in performance between "GMM-UBM with TZ-normalisation" and i-vector for the two SNR levels adopted. However, i-vector performs significantly better when factory noise has been added to the audio files. This is especially the case for the lower noise level (i.e. SNR of 15 dB). It should also be noted in this context that the background speech data used for the TZ-normalisation technique is contaminated with the same level and type of noise as the training and testing data. This is somewhat similar to the CT-norm method presented in [2]. Additionally, the experiments in this study have been based on the use of identical levels and types of noise in the training and testing data. As part of further work in this area, it is

important to evaluate the effects of noise mismatch on the performance of OSTI-SI.

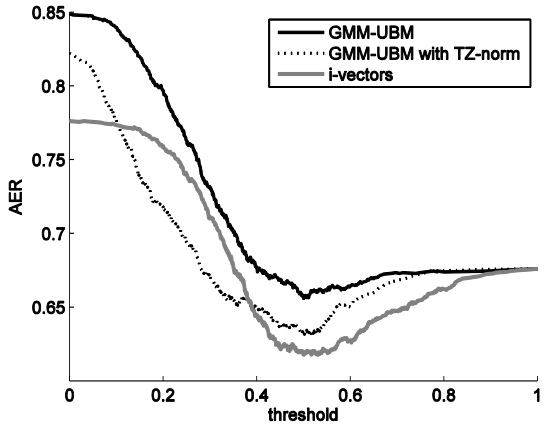

Figure 7: AER plots for speech data contaminated with factory noise (SNR =5 dB).

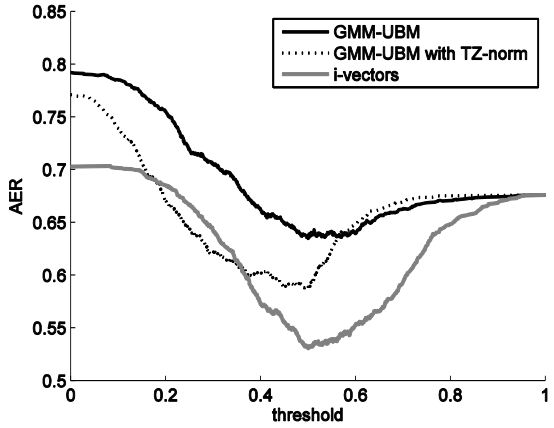

Figure 8: Experimental results for speech data moderately contaminated by factory noise (SNR = 15 dB).

In order to consider the intra-speaker variability compensation offered by i-vector, a set of experiments is conducted using the within class covariance normalization (WCCN) as described in section II.B. The result of this experimental investigation is presented in Figure 9 for the NIST telephone-quality speech data, together with the results for the same data condition presented earlier in Figure 2. As observed, the incorporation of WCCN appears to further improve the recognition performance of the i-vector technique.

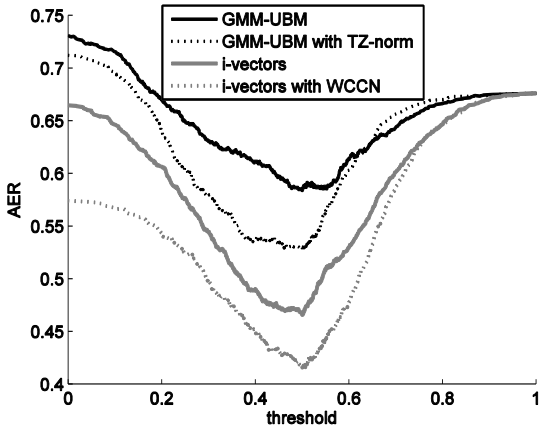

Figure 9: Results illustrating the superior performance of i-vector with WCCN in experiments based on NIST telephone quality data.

IV. Conclusion and Future Work

Overall, the experimental findings for OSTI-SI show that in comparison with the more traditional GMM-UBM approaches, the i-vector technique tends to be more robust against noise contamination of the speech data. However, the level of superiority of this approach appears to vary somewhat with the type and level of additive noise in speech.

Building on the outcomes of the investigations conducted in this study, future work should include the incorporation of the WCCN technique into the i-vector approach for experiments with various levels and types of noise contaminations. This is to examine whether or not the WCCN technique remains effective under such real-world conditions.

For high levels of noise contamination, the outcomes indicate the necessity to consider alternative or additional methods for enhancing the OSTI-SI accuracy. For example, the approach based on multi-SNR UBMs has shown promising results in [2]. A strategy that might further the accuracy of the i-vector approach, even in the presence of high levels of noise, could be that based on using multiple total variability matrices as well as multi-SNR UBMs for various signal-to-noise ratios.

References

[1] A. M. Ariyaeeinia, J. Fortuna, P. Sivakumaran *et al.*, "Verification effectiveness in open-set speaker identification, *IEE Proceedings* -" *Vision, Image and Signal Processing, ,* vol. 153, no. 5, pp. 618-624, 2006.

[2] S. Pillay, A. Ariyaeeinia, P. Sivakumaran *et al.*, "Effective speaker verification via dynamic mismatch compensation," *IET,Biometrics,* vol. 1, no. 2, pp. 130-135, 2012.

[3] N. Dehak, P. Kenny, R. Dehak *et al.*, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, no. 4, pp. 788-798, 2011.

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

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, no. 1–3, pp. 19-41, 2000.

[6] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia *et al.*, "Relative effectiveness of score normalisation methods in open-set speaker identification," *Odyssey -2004*, pp. 369-376, 2004.

[7] N. Dehak, R. Dehak, P. Kenny *et al.*, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," in 10th Annual Conference of the International Speech Communication Association (Interspeech), Brighton, United Kingdom, 2009, pp. 1559-1562.

[8] D. Garcia-Romero, and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in 12th Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy, pp 256-259, 2011.

[9] A. Kanagasundaram, D. Dean, R. Vogt *et al.*, "Weighted LDA techniques for i-vector based speaker verification," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 4781-4784.

[10] *The NIST Year 2008 Speaker Recognition Evaluation Plan*, http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf, 2008.

[11] *The NIST Year 2005 Speaker Recognition Evaluation Plan*, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.doc, 2005.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing,* vol. 10, no. 1–3, pp. 42-54, 2000.

[13] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia *et al.*, "Relative effectiveness of score normalisation methods in open-set speaker identification," in ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo Spain, 2004, pp. 369-376.

[14] A. Varga, and H. J. M. Steenken, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *Speech Communication*, pp. 247-252, 1993.

[15] A. Miró, "Robust speaker diarization for meetings," Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, http://www.xavieranguera.com/phdthesis, 2006.

[16] A. Malegaonkar, and A. Ariyaeeinia, "Performance Evaluation in Open-Set Speaker Identification," in The Third European Workshop on Biometrics and Identity Management (BioID 2011), Brandenburg, Germany, 2011, pp. 106-112.