ON THE ENHANCEMENT OF SPEAKER IDENTIFICATION ACCURACY USING WEIGHTED BILATERAL SCORING

A. Malegaonkar
Trinity Convergence Ltd,
Pride Portal, Bahirat wadi,
Pune, India
amalegaonkar@
trinityconvergence.com

A. Ariyaeeinia
University of Hertfordshire,
College Lane, Hatfield, UK
a.m.ariyaeeinia@herts.ac.uk

P. Sivakumaran
University of Hertfordshire,
College Lane, Hatfield, UK
p.sivakumaran@herts.ac.uk

J. Fortuna
Intrinsica Networks Ltd,
Meadway Technology Park,
Stevenage, UK
jfortuna@intrinsica.co.uk

Abstract - This paper presents investigations into an effective bilateral scoring method in open-set speaker identification. The approach is based on the fact that two different speakers usually are not reciprocal. A difficulty in deploying bilateral scoring is that test utterances are normally much shorter than training utterances. To tackle this problem, the proposed approach provides the final identification score based on a weighted combination of independently normalised forward and reverse scores. Based on the experimental results obtained using clean and telephone quality speech, it is shown that the proposed approach is more effective than the conventional scoring methods in open-set speaker identification.

Index Terms – Speaker identification, Bilateral scoring, Score normalisation

I. INTRODUCTION

In general, speaker identification is the process of determining the correct speaker of a given test utterance from a population of registered speakers. If this process includes the option of declaring that the test utterance does not belong to any of the registered speakers, then it is specifically referred to as open-set speaker identification. Given a set of registered speakers and a sample test utterance, this task is defined as a twofold problem [1,2,3]. Firstly, it is required to identify the speaker model in the registered set that best matches the given test utterance. This is the process of identification. Next, it is required to determine if the test utterance is actually produced by the best matched speaker or it is originated by a speaker from outside the registered set. This is the process of verification. When the speaker is not required to provide an utterance of a specific text, the task is called Open-Set, Text Independent Speaker Identification (OSTI-SI). The process of OSTI-SI is summarised in Fig. 1.

As shown in this figure, each speaker in the system is represented by a statistical model. These models are produced using the training data (cepstra) for the individual speakers. As discussed in the literature, a dominant approach for this purpose is that of Gaussian mixture modelling of the training data [4,5]. Such a model can either be obtained exclusively from the training data using the Expectation-Maximisation (E-M) algorithm [4] or can be obtained by adapting an independent background model, using the Maximum a priori (MAP) training procedure [5].

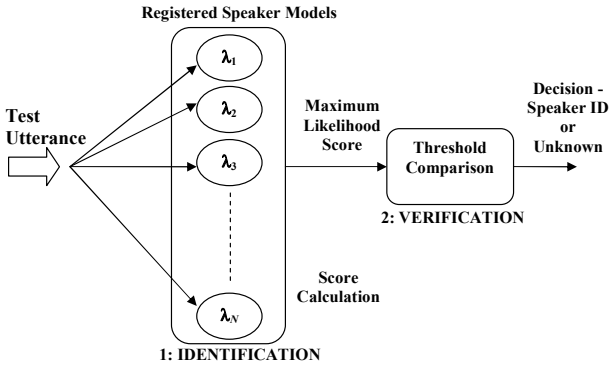

Fig.1. Process of OSTI-SI

An important factor affecting the performance of OSTI-SI in practice is the size of the population of registered speakers [3,6]. As this population grows, the confusion in discriminating amongst the voices of registered speakers is likely to increase and therefore the number of incorrect identifications is likely to increase as well. The growth in this 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 size of the population grows, the possibility of a voice originating from a non-registered speaker being very close to one of the registered speakers also increases. Thus, both processes in OSTI-SI are affected by the size of the population of registered speakers.

Undesired variations in speech characteristics due to anomalous events further complicate the task of OSTI-SI. These anomalies can have various forms ranging from variations in communication channel and environment noise to uncharacteristic sounds generated by the speaker [1,2,3]. These cause mismatches between the test utterances and the pre-stored voice patterns of registered speakers. In OSTI-SI, the second process is far more susceptible to these mismatches than the first process. This is because, in the first process, since the same test utterance is compared against all the registered models, the distortion in the test utterance is likely to be reflected similarly in all the scores. Therefore, the selection of the best matched speaker model, that is the model that yields the maximum likelihood, is unlikely to be affected. On the other hand, in the second process, this best matched score

is compared against a threshold determined a priori and without any knowledge of such distortions. Hence the second process is more likely to be affected by such contaminations. The effect of such contaminations can be alleviated, to some extent, by using score normalisation techniques [1,2,3].

The score normalisation techniques used in the previous studies are deployed in the framework of Unilateral scoring (ULS). In this framework, given a test utterance and a registered speaker model (built using some training utterances), the matching score is given in terms of a conditional probability of a speaker model generating the test utterance. In this study, the process of OSTI-SI is evaluated in the framework of more robust scoring framework of Weighted Bilateral Scoring (WBLS).

The remainder of the paper is organised as follows. Section II provides a description of the proposed weighted bilateral scoring method. Section III details the experimental investigations and presents an analysis of the results. The overall conclusions are given in Section IV.

II. WEIGHTED BILATERAL SCORING

It is reported in speaker recognition that two different speakers are usually not reciprocal [7]. That is, when the models built using speech from a speaker (speaker A) are matched against speech from another speaker (speaker B), they may not return high likelihoods whilst speech from speaker A matched against the models built using speech from speaker B giving high likelihoods [7]. This is shown in Fig. 2 which presents the results for 10 reciprocity tests, each between a pair of speakers. In each case, each of the two speakers is modelled using a single Gaussian model which is then tested against the data from the other speaker.

Such non-reciprocity may also exist (to some extent) when an utterance from a particular speaker is matched against the model built using another utterance from the same speaker. This can be observed in Fig. 3 for the tests conducted using 10 speakers. Comparing with Fig. 2, it can be said that non-reciprocity between two different speakers is more significant than the non-reciprocity between utterances originated from the same speaker.

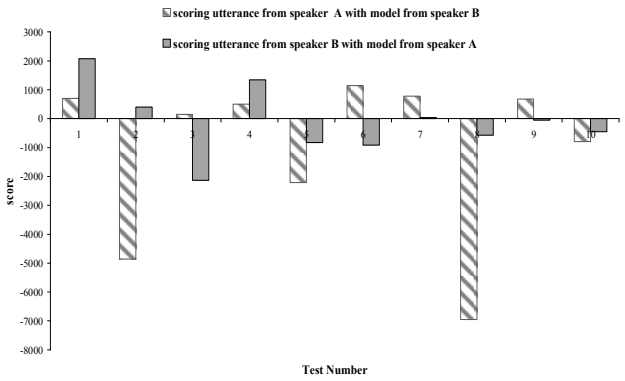

Fig. 2. Non-reciprocity between different speakers

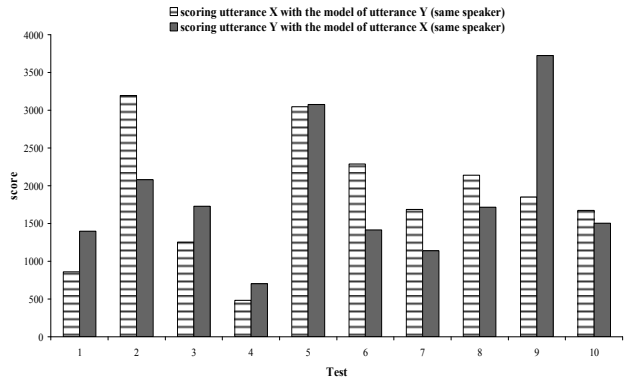

Fig. 3. Non-reciprocity between utterances from the same speaker

It is well established in the previous study in speaker verification [7] that Bilateral Scoring (BLS) is more effective than the ULS process.

Given a test utterance $\mathbf{O}_{tst}$ and a registered speaker model, $\boldsymbol{\lambda}_n^{Tr}$, built from the training utterance $\mathbf{O}_n^{Tr}$, of n^{th} registered speaker, the bilateral score is given as

$$S = p\left(\boldsymbol{\lambda}_n^{Tr} \mid \mathbf{O}_{tst}\right) \times p\left(\boldsymbol{\lambda}_{tst} \mid \mathbf{O}_n^{Tr}\right) \tag{1}$$

where $p(.)$ is a probability function, $\boldsymbol{\lambda}_{tst}$ is the model built from the test utterance $\mathbf{O}_{tst}$, and $\mathbf{O}_n^{Tr}$ is the training utterance which is used to build the model .

Equation (1) can be expanded in the Bayesian framework as

$$S = \frac{p(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_n^{Tr})\, p(\boldsymbol{\lambda}_n^{Tr})}{p(\mathbf{O}_{tst})} \times \frac{p(\mathbf{O}_n^{Tr} \mid \lambda_{tst})\, p(\boldsymbol{\lambda}_{tst})}{p(\mathbf{O}_n^{Tr})} \tag{2}$$

It should be noted that $p(\boldsymbol{\lambda}_n^{Tr})$ is the same for all speaker models and can therefore be ignored. For the same reason, $p(\boldsymbol{\lambda}_{tst})$ can be ignored in the calculation of bilateral scores. Based on the above, equation (2) can be expressed in the log domain as

$$\rho = \log(S) = L(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_n^{Tr}) - L(\mathbf{O}_{tst}) + L(\mathbf{O}_n^{Tr} \mid \boldsymbol{\lambda}_{tst}) - L(\mathbf{O}_n^{Tr}) \tag{3}$$

where $L(.)$ is the log likelihood function.

The terms $L(\mathbf{O}_{tst})$ and $L(\mathbf{O}_n^{Tr})$ can be approximated through the use of an appropriate background model, $\boldsymbol{\lambda}_{BG}$, and hence replaced by $L(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_{BG})$ and $L(\mathbf{O}_n^{Tr} \mid \boldsymbol{\lambda}_{BG})$ respectively.

One of the difficulties in deploying bilateral scoring in OSTI-SI is that the duration of the test utterance, $\mathbf{O}_{tst}$, can be rather short (e.g. 1-5 seconds). Obtaining a GMM of a reasonable order (e.g. 32) using the E-M algorithm from such a sparse data may not always be possible. A better approach in such conditions is to use adapted GMMs [2,3,5,8]. The adaptation procedure is shown to be robust

against sparse data conditions and hence $\boldsymbol{\lambda}_{tst}$ can be obtained in a more reliable manner.

In practice, the training data is expected to be of longer duration and higher quality than the testing data. Hence, in bilateral scoring, the speaker model built from the training data can be considered more reliable than that built from the test data. Therefore, the forward log-likelihood scores (i.e. $L(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_n^{Tr}) - L(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_{BG})$) are expected to be more reliable than the reverse log-likelihood scores (i.e. $L(\mathbf{O}_n^{Tr} \mid \boldsymbol{\lambda}_{tst}) - L(\mathbf{O}_n^{Tr} \mid \boldsymbol{\lambda}_{BG})$). Consequently, it is envisaged that emphasising the forward scores relative to reverse scores can be beneficial. In this study, this technique is referred to as Weighted Bilateral Scoring (WBLS). In this technique, the forward and reverse scores are weighted and the resultant score is given as.

$$\begin{aligned}\rho_w = w \times \left\{ L(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_n^{Tr}) - L(\mathbf{O}_{tst} \mid \boldsymbol{\lambda}_{\mathrm{BG}}) \right\} \\ + (1-w) \times \left\{ L(\mathbf{O}_n^{Tr} \mid \boldsymbol{\lambda}_{tst}) - L(\mathbf{O}_n^{Tr} \mid \boldsymbol{\lambda}_{\mathrm{BG}}) \right\}\end{aligned} \quad (4)$$

where, w is the weight used for fusing the forward and backward scores.

$\boldsymbol{\lambda}_{BG}$, in this case, can be based on any of the established techniques of score normalisation. As discussed in [2,3], it is common to consider the performance of adapted GMMs in conjunction with the World Model Normalisation (WMN) as the baseline performance. In the same study, the technique of Unconstrained Cohort Normalisation (UCN) is shown to be quite promising for reducing the error rates in the verification process of OSTI-SI. The scope of that study is limited to ULS. In this study, the extent of effectiveness of UCN for OSTI-SI is investigated in the context of bilateral and weighted bilateral scoring.

In the case of UCN, the terms involving background model scoring, in (4), are given as

$$L(\mathbf{O}_x \mid \boldsymbol{\lambda}_{\mathrm{BG}}) = \log\left(\prod_{c=1}^{C} p\left(\mathbf{O}_x \mid \boldsymbol{\lambda}_{\phi(c)}\right) \right)^{1/C} \quad (5)$$

where, $\mathbf{O}_x$ is the speech data used in the modelling process (i.e. $\mathbf{O}_{tst}$ or $\mathbf{O}_n^{Tr}$), *C* is the cohort size of background speaker models, *p* is a likelihood function and $\boldsymbol{\lambda}_{\phi(c)}$ are *C* models from the background set which yield the highest *C* likelihood scores for $\mathbf{O}_x$.

The effectiveness of UCN in the framework of weighted bilateral scoring is investigated in the next section.

III. EXPERIMENTAL INVESTIGATION

The aim of the experiments presented in this study is to investigate the effectiveness of WBLS for OSTI-SI, in relation that of ULS and BLS. The scope of the study also includes evaluating the performance of UCN in the bilateral scoring framework. This choice of score normalisation is based a previous study reporting its effectiveness [2,3].

The first part of the investigation is concerned with the relative performance of OSTI-SI for short and long test utterances captured in clean and also telephonic conditions. The approximate lengths of the short and long utterance are 3 s and 10 s respectively. The second part of the investigation involves evaluating the performance of OSTI-SI in varied telephonic conditions and with test utterances of unconstrained lengths.

A. Speech Data

The speech data used for the first part of this investigation is based on the TIMIT and the NTIMIT databases. NTIMIT is the telephonic version of the TIMIT database. In these databases, there are 10 short utterances for every speaker. Out of the 10 utterances, 7 utterances are concatenated together and used for training the speaker models. The remaining 3 utterances are used for the testing purpose. This constitutes three short test utterances for every speaker. For conducting the tests using long utterances, the selected three short utterances are concatenated together to form a single test utterance for every speaker. The same experimental configuration is used for the datasets obtained using TIMIT and NTIMIT databases. This configuration is summarised in Table I. It should be noted that, in this case, the datasets are gender-balanced (The number of male and female speakers are approximately the same).

TABLE I
CONFIGURATION OF THE DATASETS BASED ON TIMIT AND NTIMIT DATABASES

	Short	**Long**
Number of registered speakers	100	100
Number of tests from registered speakers	300	100
Number of non-registered Speakers	80	80
Number of tests from non-registered speakers	240	80
Number of speakers for the world model	100	100
Length of the data for the world model	1 hr	1hr

For the second part of this study, the dataset is based on the NIST-Speaker Recognition Evaluation (SRE) 2003 database. The configuration of this dataset is based on the protocol detailed in [1,2,3] and is given in Table II. The duration of the test segments in this dataset are between 3 and 60 seconds.

TABLE II
CONFIGURATION OF THE DATASET BASED ON THE NIST SRE 2003 DATABASE

Number of registered speakers	142
Number of tests from registered speakers	1293
Number of non-registered Speakers	141
Number of tests from non-registered speakers	1408
Number of speakers for the world model	100
Length of the data for the world model	8hrs

B. Feature Representation

Each speech frame of 20 ms duration is subjected to pre-emphasis and then analysed to extract a p^{th} order linear predictive coding-derived cepstral (LPCC) feature vector at a rate of 10 ms. The value of p is chosen as 20 for the datasets obtained using TIMIT and NTIMIT databases and 12 for the dataset obtained using the NIST SRE 2003 database. The static features are mean normalised. The first derivative parameters are also extracted through a polynomial fit over 15 frames. These parameters are appended to the static features.

C. Speaker Representation

In this work, the experiments are performed using adapted Gaussian Mixture models as they are established as the most effective speaker representation for OSTI-SI [2]. The adapted models in this study have 2048 components. For the adaptation purpose, a gender independent world model is first obtained by pooling together two gender dependant world models. The adapted models are then obtained using a single step Bayesian adaptation procedure as given in [5]. For background modelling in UCN, the speakers in the registered set are used.

D. Testing Procedure

For each test utterance, the log likelihood scores (both bilateral and unilateral) are obtained using the fast scoring procedure given in [2,3]. As recommended in that study, only the top scoring mixture is used in each case.
The performance of OSTI-SI is given in terms of equal error rate (EER) and identification error rate (IER). The baseline performance is based on incorporating WMN in the scoring procedure. The overall performance is obtained by applying UCN in each individual case.

E. Results and Discussions

The first set of experiments in this study investigates the effectiveness of BLS and WBLS for short test utterances (around 2-3 seconds). As discussed above, UCN is adopted for the purpose of this study. The experiments are carried out for clean and telephonic audio conditions. The results are given in Table III and Table IV respectively. In these tables, ULS forward and ULS reverse refer to the unilateral scoring procedures involving forward and reverse scoring respectively.

TABLE III
RESULTS FOR SHORT UTTERANCES FROM TIMIT

	IER(%)	EER(%)
ULS forward	8.3	33.3
ULS forward-UCN	8.3	12.9
ULS reverse	10.0	34.1
ULS reverse-UCN	10.0	14.8
BLS	7.6	32.5
BLS-UCN	7.6	12.3
WBLS	7.6	32.3
WBLS-UCN	7.6	10.8

TABLE IV
RESULTS FOR SHORT UTTERANCES FROM NTIMIT

	IER(%)	EER(%)
ULS forward	66.6	40.0
ULS forward-UCN	66.6	32.2
ULS reverse	67.0	41.7
ULS reverse-UCN	67.0	34.2
BLS	66.3	39.2
BLS-UCN	66.3	27.5
WBLS	65.3	38.1
WBLS-UCN	65.3	23.7

It can be observed from these tables that the performance of OSTI-SI significantly deteriorates in the telephonic condition. The technique of unilateral forward scoring can be observed to perform better than unilateral reverse scoring. As discussed in Section II, the main reason for this is the higher reliability of speaker models in the forward scoring procedure.
BLS and WBLS, without UCN, are observed to perform marginally better than ULS for both TIMIT and NTIMIT. The advantage of using UCN is seen to be rather considerable in the case of clean data conditions. In this case, BLS and WBLS are found to perform significantly better than ULS. It is also noted that in all cases, WBLS appears to outperform the other scoring procedures. Particularly, improvement is observed by lowering of EER. For clean acoustic condition, WBLS-UCN is seen to outperform ULS forward-UCN, in terms of EER, by about 17%. This performance difference is observed to increase to about 26% (in favour of WBLS-UCN) in the degraded audio condition. Table IV also shows that, in terms of IER, WBLS again performs better than ULS, but the performance difference in this case is marginal.
In this study, a cohort size between 2 and 7 is found to lead to the best results in all the cases considered, and a weight in the range 0.65 to 0.85 is observed to be most beneficial for WBLS.
The next set of experiments in this study investigates the effectiveness of the proposed approach for the adopted long test utterances. Similar to the previous set of experiments, the performance of OSTI-SI is observed in the clean and telephonic acoustic conditions. The results are given in Table V and VI respectively.

TABLE V
RESULTS FOR LONG UTTERANCES FROM TIMIT

	IER(%)	EER(%)
ULS forward	1	30.0
ULS forward-UCN	1	5.0
ULS reverse	1	31.2
ULS reverse-UCN	1	13.8
BLS	1	29.3
BLS-UCN	1	3.9
WBLS	1	28.4
WBLS-UCN	1	3.7

TABLE VI
RESULTS FOR LONG UTTERANCES FROM NTIMIT

	IER (%)	EER (%)
ULS forward	38	31.2
ULS forward-UCN	38	27.5
ULS reverse	37	33.7
ULS reverse-UCN	37	29.0
BLS	33	32.4
BLS-UCN	33	25
WBLS	33	31.2
WBLS-UCN	33	21.3

Comparing these results with those in Table III and IV, it can be observed that, in general, significantly better performance is obtained with long test utterances than with short test utterances. For both the audio conditions and all the techniques considered, the use of UCN is observed to improve the performance of verification process in OSTI-SI. This is more significant in the case of clean audio condition. Similar to the case in the previous set of experiments (short test utterances), it is observed that WBLS-UCN is more effective than forward ULS-UCN as well as BLS-UCN.

The final set of experiments in this study is aimed at observing the performance of the proposed approach using a dataset derived from the NIST SRE 2003 database. The results for this dataset are given in Table VII.

TABLE VII
RESULTS FOR THE SPEECH DATA OBTAINED FROM THE NIST SRE 2003 DATABASE

	IER (%)	EER (%)
ULS forward	32.7	21.6
ULS forward-UCN	32.7	19.1
ULS reverse	58.5	29.8
ULS reverse-UCN	58.5	26.7
BLS	32.7	21.1
BLS-UCN	32.7	18.1
WBLS	32.5	20.7
WBLS-UCN	32.5	15.2

It can be observed from this table that, the proposed approach yields the best performance. It is mainly beneficial in improving the performance of the verification process in OSTI-SI. In this case, the performance improvement achieved with WBLS-UCN over ULS-UCN is in excess of 20%.

IV. CONCLUSIONS

A weighted bilateral scoring method for open-set speaker identification has been proposed and investigated. The approach, which involves exploiting the fact that two different speakers are not usually reciprocal, produces the similarity score as a weighted combination of forward and reverse scores. The experiments with clean and telephone quality speech have shown that, in practice, forward scores are normally more accurate than reverse scores. This is due to the fact that the speech samples obtained for the enrolment of speakers are usually longer in duration than the speech samples taken in the test phase. As a result, the speaker models built using the enrolment speech material are more reliable than those built using the test tokens. The proposed approach deals with this imbalance in the quality of speaker models by appropriately weighting the forward and reverse similarity scores before combining them. Based on the experimental results it is shown that the identification error rate (IER) with weighted bilateral scoring is consistently lower than that obtained with conventional unilateral scoring. Moreover, it is demonstrated experimentally that the equal error rate (EER) in the second stage of open-set speaker identification can be considerably reduced by using the proposed weighted bilateral scoring together with unconstrained score normalisation.

V. REFERENCES

[1] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalisation methods in open-set speaker identification," in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2004),* pp. 369-376, 2004.

[2] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Open-set speaker identification using adapted Gaussian mixture models," in *Interspeech, Lisbon*, Portugal, 2005

[3] A.M. Ariyaeeinia, J. Fortuna, P. Sivakumaran and A. Malegaonkar, "Verification Effectiveness in Open-Set Speaker Identification", *IEE Proceedings Vision, Image and Signal Processing*, Vol. 153, Issue 5, pp. 618-624, Oct 2006.

[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, 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, pp. 19-41, 2000.

[6] U. Chaudhari, J. Navratil, G. Ramaswamy, and S. Maes, "Very large population text-independent speaker identification using transformation enhanced multi-grained models," in *ICASSP, 2001*.

[7] E. S. Parris and M. Carey, "Multilateral techniques for speaker recognition," *presented at ICSLP*, *Sydney*, Australia, 1998

[8] A. Malegaonkar, A.M. Ariyaeeinia and P. Sivakumaran, "Efficient Speaker Change Detection Using Adapted Gaussian Mixture Models", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, Issue 6, pp. 1859-1869, August 2007.