INTERSPEECH
2009 BRIGHTON

Open-Set Speaker Identification under Mismatch Conditions

*S. G. Pillay, A. Ariyaeeinia, P. Sivakumaran and M. Pawlewski**

University of Hertfordshire, Hatfield, UK
*BT Labs, Ipswich, UK
{s.g.pillay, a.m.ariyaeeinia, apb}@herts.ac.uk, *mark.pawlewski@bt.com

10.21437/Interspeech.2009-395

Abstract

This paper presents investigations into the performance of open-set, text-independent speaker identification (OSTI-SI) under mismatched data conditions. The scope of the study includes attempts to reduce the adverse effects of such conditions through the introduction of a modified parallel model combination (PMC) method together with condition-adjusted T-Norm (CT-Norm) into the OSTI-SI framework. The experiments are conducted using examples of real world noise. Based on the outcomes, it is demonstrated that the above approach can lead to considerable improvements in the accuracy of open-set speaker identification operating under severely mismatched data conditions. The paper details the realisation of the modified PMC method and CT-Norm in the context of OSTI-SI, presents the experimental investigations and provides an analysis of the results.

1. Introduction

In general, the problem of automatic speaker identification can be defined as one of determining the speaker of a given test utterance, from a population of registered speakers [1]. If the process includes the option of declaring that the test utterance does not belong to any of the registered speakers, it is termed open-set speaker identification. Otherwise, it is a closed-set identification process [1-3]. In principle, the process of open-set speaker identification consists of two successive stages of identification and verification. In other words, first, it is required to identify the speaker model in the set, which best matches the given test utterance. Then, it must be verified whether the test utterance has actually been spoken by the speaker associated with the best-matched model, or by some unknown speaker outside the registered set [1]. When there are no constraints on the text content of test utterances, the process is referred to as open-set, text-independent speaker identification (OSTI-SI). This is the most challenging class of speaker recognition with applications in various areas including document indexation, surveillance, and authorisation control in smart environments.

A factor adversely affecting the accuracy of OSTI-SI in practice is that of variations in speech characteristics [1, 2]. Such variations occur due to various causes such as environmental noise, channel effects, or uncharacteristic sounds by the speakers (e.g. lip smacks) [1, 3]. The net result is a mismatch between the corresponding test and reference material for the same speaker, which in turn reduces the accuracy of OSTI-SI.

To date, there has been considerable research into speaker recognition under mismatch conditions [3-5]. A widely used approach for tackling this problem with a considerable degree of success is that of score normalisation. The approach is based on obtaining a normalisation factor using the match score(s) computed for the test utterance against a set of background (competing) models or a single universal background model [3, 6]. In general, however, the effectiveness of score normalisation reduces considerably when the data mismatch, resulting from noise contamination in the test material, becomes significant [4]. To tackle this problem, a modified form of T-Norm, termed CT-Norm (condition-adjusted T-Norm) has recently been proposed by the authors and investigated in the context of speaker verification (SV) [4]. The technique, which is based on an efficient parallel model combination (PMC) approach, has been shown to significantly reduce the adverse effects of data mismatch on SV systems.

However, as indicated earlier, the problem in the second stage of OSTI-SI is more challenging than that of the standard speaker verification [1, 2, 7]. This is due to the fact that the requirement in the second stage of OSTI-SI is to discriminate each out-of-set speaker from its best matched speaker in the registered set. Therefore, it may not be possible to fully predict the effectiveness of CT-Norm in this case, based on the results obtained for SV [4]. Moreover, the benefits of using the computationally efficient PMC GMM-UBM approach for speaker identification have not been previously investigated. The aim of this paper is to complement the previous study by investigating the effectiveness of the efficient PMC GMM-UBM approach and CT-Norm in the context of open-set speaker identification.

The remainder of the paper is organised in the following manner. The next section describes the approach to using the modified (efficient) PMC GMM-UBM and CT-Norm for OSTI-SI. The experimental investigations together with an analysis of the results are presented in Section 3. The overall conclusions are given in Section 4.

2. Modified PMC GMM-UBM for OSTI-SI

The study in [5] has shown that an approach to reducing the effects of data mismatch in speaker verification is that based on the use of the PMC (data-driven Parallel Model Combination) technique. The outcomes of that study provide a clear indication of the effectiveness of PMC in speaker verification based on decoupled modelling. In a more recent study [4], the authors have introduced a modified PMC method for speaker verification based on GMM-UBM [6] and GMM-SVM [8]. The fundamental problem in using PMC directly with the said coupled modelling approaches is that of computational cost. To be more specific, in the case of GMM-UBM (or GMM-SVM), the direct use of PMC involves rebuilding a UBM (with appropriately degraded speech material) as well as the adaptation of the new UBM using the degraded version of the training utterances for the target speaker. Repeating the process of rebuilding a new UBM in each test trial is computationally very expensive and inefficient.

The modified PMC approach introduced in [4] tackles this problem by using a fixed UBM for all test trials. Such a fixed UBM is built using the appropriate speech data available for this purpose in the training phase.

According to the study in [4], despite its enhanced efficiency, the use of the modified PMC with the GMM-SVM approach still results in an undesirably high level of computational cost. This is mainly due to the specific characteristics of this SVM-based approach which make the incorporation of the modified PMC unsuitable for most practical applications [4]. For this reason, the GMM-SVM [8] classification method is not considered in the present study.

Figure 1 illustrates the use of the modified PMC approach with GMM-UBM for OSTI-SI. As shown in this figure, an estimate of the test utterance degradation is used to contaminate the training utterances of the registered speakers. The noise-adjusted registered speaker models are then built by appropriately adapting the fixed (original) UBM using a modified MAP estimation (hereafter referred to as *m*MAP) [2, 6]. Once the new models are obtained, the test utterance is matched against all the registered speaker models and the model that yields the largest score is retained. This process is based on the fast scoring procedure using the top five scoring UBM mixtures identified for each test feature vector [6]. As indicated in Figure 1, the score for the speaker model selected as above is then subjected to normalisation using T-Norm.

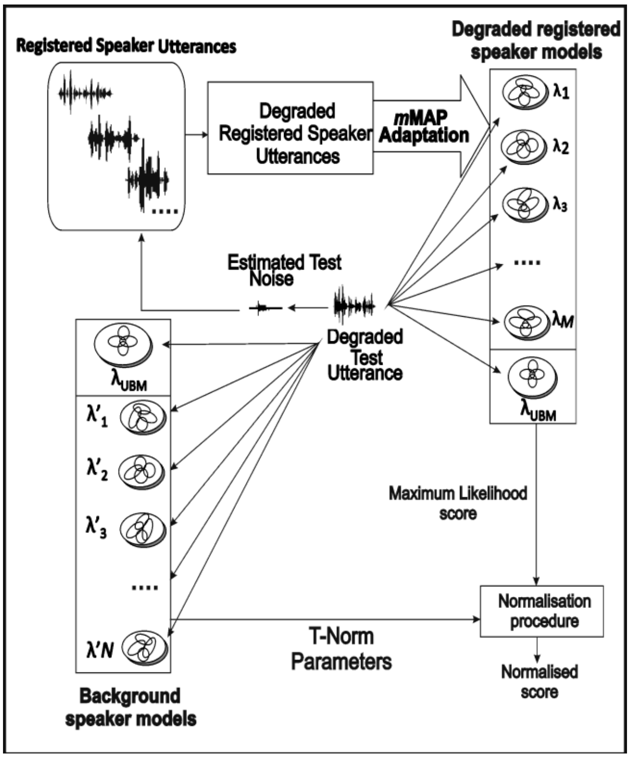

Figure 1*: OSTI-SI based on the modified PMC GMM-UBM approach.*

It has, however, been observed from the investigations in [4] that the effectiveness of T-Norm is rather limited when the degradation in the test utterance is considerably higher than that in the training material. To overcome this problem, the concept of condition adjusted T-Norm (CT-Norm), which has recently been introduced by the authors and investigated in the context of SV [4], is incorporated in the OSTI-SI framework. As shown in Figure 2, the method involves adjusting the noise contamination of background speaker utterances (and hence their models), in accordance with the estimated test utterance degradation. The determination of the normalisation parameters is then based on these contaminated background speaker models.

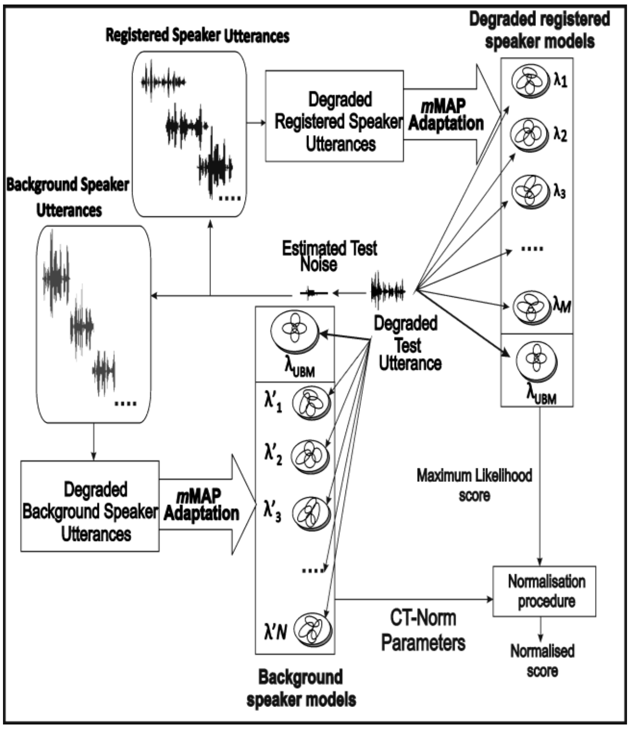

Figure 2*: OSTI-SI based on the modified PMC GMM-UBM approach with CT-Norm.*

3. Experimental investigation

3.1. Speech data

The speech dataset used for the purpose of the experimental investigations is extracted from the TIMIT database. 100 registered speakers and 80 unknown speakers are used, each having 10 utterances. Utterances from 100 speakers, other than the ones registered or considered as unknown speakers, are used for training a UBM. As in [4], it should be noted that the speaker set used for UBM and the sets of registered and unknown speakers are all gender-balanced.

In order to facilitate the experimental investigations, in each test trial, the implementation of CT-Norm (or T-Norm where appropriate) is based on the use of the training utterances from the cohort of speakers available within the set of registered users (i.e. 99 speakers on each occasion). It should also be pointed out that the main reason for using the TIMIT database in this study is that it offers the flexibility required for investigating the effectiveness of the modified PMC GMM-UBM with CT-Norm approach under controlled conditions.

3.2. Feature extraction

In this study, each speech frame of 20 ms duration is subjected to a pre-emphasis process and then used to obtain a 20th order mean-subtracted, linear predictive coding-derived cepstral vector (LPCC), extracted at a rate of 10 ms. Delta parameters are also calculated and appended to the static features [1].

3.3. Experimental setup, results and discussions

3.3.1. Effects of Mismatched Data Conditions

The aim of the first set of experiments is to determine the effectiveness of GMM-UBM for OSTI-SI in the absence of information about the relative noise conditions in the test and training phases. For this purpose, clean training data is used during the modelling process while degraded data is used in the test phase [4]. Three examples of real-world noise (i.e. car noise, office noise, and factory noise), obtained from the NOISEX 92 [9] and Piper [10] databases, are used to degrade the test data; achieving SNRs of 15dB, 10dB and 5dB.

Table 1 presents the results in terms of identification error rate (IER) and open set identification equal error rate (OSI-EER) with a 95% confidence interval. It is observed that for all the real world noise considered, there is a substantial increase in error rates (OSI-EERs and IERs) with decreasing SNR. This is particularly significant for the IERs where a difference in performance of over 50% is observed for data SNRs of 10dB and 5dB. To further illustrate the effects of mismatch conditions on the accuracy of OSTI-SI, the results in Table 1 should be compared with those in Table 2 which are obtained under clean matched data conditions. These results clearly outline the negative impacts on both the OSI-EERs and IERs, which occur from varying levels of noise degradation between the training and testing data. It is also noted that, unlike in match conditions, the benefits of T-Norm are very limited.

OSI-EER (%)				
GMM-UBM				
Noise		*15dB*	*10dB*	*5dB*
Car	Clean UBM	20.50 ±1.94	26.50 ±2.29	31.13 ±3.29
	T-Norm	17.25 ±1.82	24.38 ±2.23	30.38 ±3.24
IER(%)		14.20	26.00	59.60
Office	Clean UBM	19.38 ±1.91	23.88 ±2.25	31.75 ±3.57
	T-Norm	17.50 ±1.84	22.75 ±2.21	31.13 ±3.55
IER(%)		15.20	28.00	66.00
Factory	Clean UBM	20.37 ±1.94	23.75 ±2.16	37.5 ±3.79
	T-Norm	18.25 ±1.85	22.37 ±2.12	33.5 ±3.68
IER(%)		13.40	22.60	67.200

Table 1*: Accuracy of OSTI-SI under mismatch conditions.*

OSI-EER (%)	
GMM-UBM	
UBM	13.50 ±0.62
T-Norm	8.00±0.56
IER (%)	5.20

Table 2*: Performance of OSTI-SI under clean match conditions.*

3.3.2. Performance of the condition adjusted approach

To examine the relative effectiveness of the modified PMC-GMM-UBM with CT-Norm approach for OSTI-SI, a set of experimental investigations is conducted, using the setup described in Section 2. As before, the same set of real world noise is used to degrade the test data. For the purpose of modified PMC, a 200 ms segment of noise is used as the estimation of test utterance contamination. The results for this part of the experimental investigations are presented in Table 3.

OSI-EER (%)				
modified PMC GMM-UBM				
Noise		*15dB*	*10dB*	*5dB*
Car	Clean UBM	23.88 ±1.96	26.75 ±2.04	31.25 ±2.23
	CT-Norm	12.62 ±1.53	17.87 ±1.76	22.00 ±1.99
IER(%)		5.40	5.80	13.60
Office	Clean UBM	20.13 ±1.86	26.38 ±2.11	37.50 ±2.63
	CT-Norm	13.00 ±1.56	18.25 ±1.85	23.00 ±2.29
IER(%)		6.60	12.60	32.20
Factory	Clean UBM	22.50 ±1.91	27.63 ±2.06	36.50 ±2.32
	CT-Norm	13.63 ±1.58	15.75 ±1.68	23.00 ±2.03
IER(%)		4.80	5.80	14.00

Table 3*: Performance the condition adjusted approach.*

There are a number of interesting observations which can be made from the above table. Firstly, as expected, it is observed that the use of the modified PMC GMM-UBM on its own does not have any considerable benefits on the accuracy in the second stage of OSTI-SI. It is, however, seen that the said process is considerably beneficial to the accuracy in the first stage, leading to significant improvements in IER for all types of noise considered. For instance, when the test data quality is reduced to 5 dB using factory noise, the improvement achieved in IER (relative to that in Table 1) is in excess of 79%. In addition, it is observed that CT-Norm is considerably more effective than T-norm in reducing OSI-EER. Considering all type of noise and degradation levels in this study, the average improvement achieved in OSI-EER relative to the best results in Table 1 is about 25%. The relative improvements offered by the CT-Norm approach under mismatch conditions are further illustrated through the DET plots in Figure 3. In all cases, the SNR for the test data is 5dB.

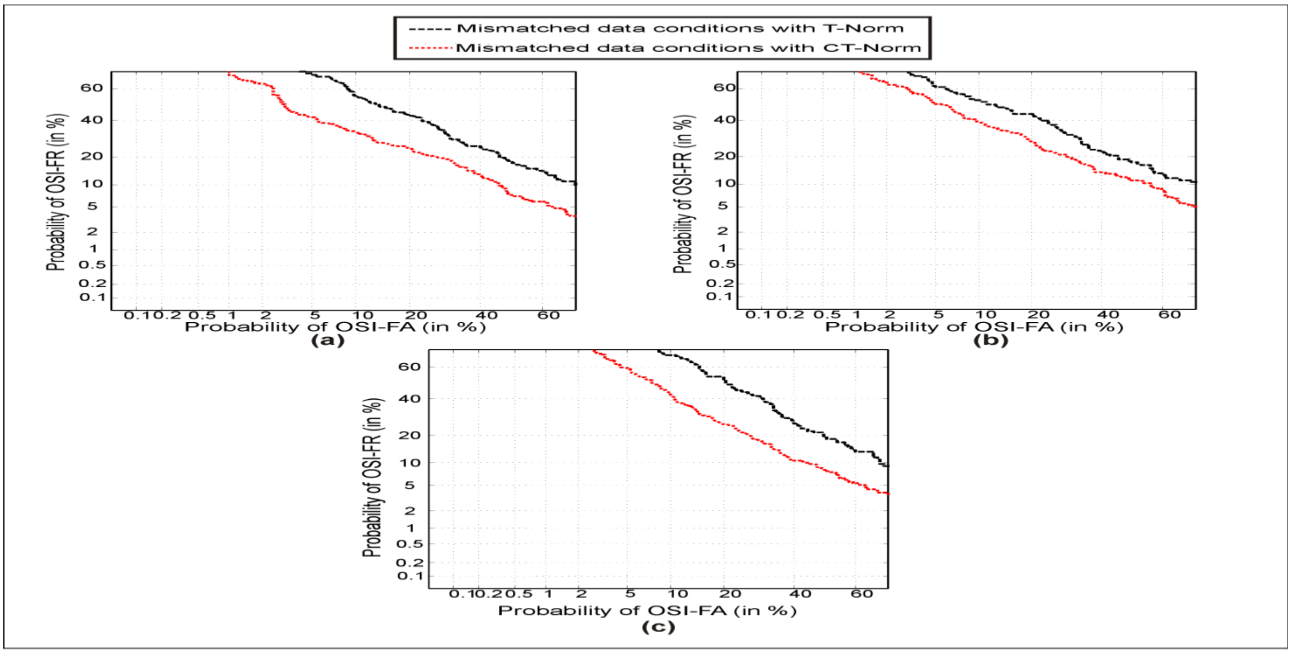

Figure 3: *Relative verification effectiveness offered by the use of CT-norm with the modified PMC GMM-UBM approach in mismatched data conditions using (a) car noise (b) office noise (c) factory noise.*

It is also important to compare the results in Table 1 and 3 with the corresponding results obtained under the same experimental conditions for speaker verification [4]. Such a comparison can clearly show that the adverse effects of mismatch data conditions are more significant in the second stage of OSTI-SI than in standard SV. With reference to [4], it also appears that the proposed method is more effective in standard speaker verification than in the second stage of OSTI-SI. These further highlight the additional challenges in the second stage of OSTI-SI.

4. Conclusions

An investigation into the relative effectiveness of the modified parallel model combination (PMC) GMM-UBM with condition adjusted T-Norm (CT-Norm) approach for OSTI-SI systems has been presented. It has been shown that the performance of OSTI-SI is severely affected when the level of degradation in the test material is different from that in the training utterances. Based on the outcomes of the experimental investigations, it is shown that in these adverse scenarios, the modified PMC GMM-UBM approach can significantly improve the accuracy of the first stage of the OSTI-SI process (up to 79% for severely degraded data conditions). It is also shown that that the use of CT-Norm with the said approach is of considerable benefit to the verification stage. In this case, the average accuracy improvement relative to conventional GMM-UBM is found to be around 25%.

An important aspect of the projected study is that of further enhancing the computational efficiency of this approach. This is because the technique can become computationally expensive as the registered speaker population grows.

5. References

[1] Ariyaeeinia, A., Fortuna, J., Sivakumaran, P. and Malegaonkar, A., "Verification effectiveness in open-set speaker identification", IEE Proceedings Vision, Image and Signal Processing, vol. 153, 618-624, 2006.

[2] Fortuna, J., Sivakumaran, P., Ariyaeeinia, A. and Malegaonkar, A., "Relative effectiveness of score normalization methods in open-set speaker identification," in Proc. Odyssey 2004 Speaker and Language Recognition, 369-376, 2004.

[3] Auckenthaler, R., Carey, M. and Thomas, H. L.,"Score normalization for text-independent speaker verification systems," Digital Signal Processing, vol. 10, 42-54, 2000.

[4] Pillay, S., Ariyaeeinia, A., Pawlewski, M. and Sivakumaran, P., "Speaker verification under mismatched data conditions," IET Signal Processing, Special Issue on Biometric Recognition, In press, 2009.

[5] Bellot, O., Matrouf, D., Merlin, T. and Bonastre, J. F., "Additive and convolutional noises compensation in speaker recognition," Proc. International Conference on Spoken Language Processing (ICSLP), 799-802, 2000

[6] Reynolds, D. A., Quatieri, T. and Dunn, R., "Speaker Verification using adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, 19-41, 2000.

[7] Malegaonkar, A., Ariyaeeinia, A., Sivakumaran, P. and Fortuna, J., "On the enhancement of speaker identification accuracy using weighted bilateral scoring," Proc Security Technology (ICCST 2008), 254-258, 2008.

[8] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., "Support Vector Machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, vol. 13, 308-311,2006.

[9] Varga, A., Steeneken H. J. M., Tornlinson, M. and Jones, D., "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Speech Research Unit, Defense Research Agency, 1992.

[10] Sivakumaran, P., "Robust Text Dependent Speaker Verification," PhD Thesis, University of Hertfordshire, 1998.