ANALYSIS AND COMPARISON OF SCORE NORMALISATION METHODS FOR TEXT-DEPENDENT SPEAKER VERIFICATION

A. M. Ariyaeeinia and P. Sivakumaran
University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK
A.M.Ariyaeeinia@herts.ac.uk, P.Sivakumaran@herts.ac.uk

ABSTRACT

This paper presents an investigation into the relative effectiveness of various score normalisation methods for speaker verification. The study provides a thorough analysis of different approaches for normalising verification scores, and comparatively examines these under identical experimental conditions. The experiments are based on the use of subsets of the Brent (telephone quality) speech database, consisting of repetitions of isolated digit utterances zero to nine spoken by native English speakers. Based on the experimental results it is demonstrated that amongst the considered methods, a particular form of the cohort normalisation method provides the best performance in terms of the verification accuracy. The paper discusses details of the experimental study and presents an analysis of the results.

1. INTRODUCTION

One of the main factors adversely affecting the performance of text-dependent speaker verification in practice is that of undesired variations in speech characteristics due to anomalous events. These anomalies can have different forms ranging from environmental and transmission channel noise to uncharacteristic speech sounds from speakers. The resultant variations in speech cause a mismatch between the corresponding test and reference patterns which in turn can lead to a significant reduction in the verification accuracy. Due to the absence of accurate information about the existence, level and nature of variations in speech characteristics in practice, it has been proposed to introduce robustness into the verification operation through an appropriate normalisation of the verification scores [1]-[4]. Although a number of methods have already been developed for this purpose [2]-[4], these have been mainly examined in independent studies and their relative effectiveness has not previously been thoroughly investigated.

This paper presents an analysis of various score normalisation methods, and details a comparative evaluation of the effectiveness of these for robust speaker verification. For the purpose of this evaluation, different normalisation methods are employed in experiments conducted under identical conditions. By drawing impostors from within and without the set of registered speakers, attempts are also made to investigate the effect of this factor on the performance of the considered normalisation techniques.

2. SCORE NORMALISATION METHODS

The use of score normalisation in speaker verification has been a direct result of the probabilistic modelling of speakers [5],[6]. By adopting this modelling method and using Bayes theorem, the verification score can be expressed as [3],[6]

$$S_i = \log \frac{p(\mathbf{O}|\lambda_i)}{p(\mathbf{O}|\lambda)}, \qquad (1)$$

where $p(\mathbf{O}|\lambda_i)$ is the likelihood of the observed feature vector sequence, $\mathbf{O}$, for the target speaker i (with the reference model λ_i), and $p(\mathbf{O}|\lambda)$ is the likelihood for any speaker. This latter likelihood can be viewed as a means of normalising the likelihood for the target speaker *i*. A modified form of (1), which is normally considered for score normalisation, is based on replacing $p(\mathbf{O}|\lambda)$ with the average of densities for all speakers other than the target speaker [6]:

$$S_i = \log \frac{p(\mathbf{O}|\lambda_i)}{p(\mathbf{O}|\lambda \neq \lambda_i)}. \qquad (2)$$

The cohort normalisation method proposed in [2] is based on approximating $p(\mathbf{O}|\lambda \neq \lambda_i)$ using a cohort of speakers whose models are most competitive with the target model. The approach involves selecting the competing speakers based on the closeness of their models to the model of the target speaker. The advantage of this method is that if the existence of anomalous events in the test utterance causes a speaker's score against his (her) own model to degrade, then the scores obtained using the same test utterance against the selected competing models may also be affected in the same way. As a result, the normalised score may remain relatively unaffected. The technique may therefore be expected to help reduce false rejection. A main

drawback of this method is that it provides the possibility of a test utterance produced by an impostor being almost equally dissimilar from the target model and the competing models. In such cases, whose frequency of occurrence depends on the closeness of the competing models to the target model, the normalised score may become large enough to lead to the acceptance of the impostor.

A method for tackling the above problem is to select the competing speaker models based on their closeness to the given test utterance [4],[7]. It can be argued that with this method, when the test utterance is produced by the true speaker, the competing models will be reasonably close to the target model. Therefore, the method can be expected to be almost as effective as the previous approach. However, when the test utterance is produced by an impostor, the selected competing models will be close to the test utterance but not necessarily to the target model. As a result, for a fixed verification threshold, the technique is capable of reducing the possibilities of both false acceptance and false rejection. Since this cohort-based approach allows the selection of competing speaker models in each test trial to depend on their relative scores on that occasion, it is referred to as unconstrained cohort in this paper.

Another approach for score normalisation involves using utterances from a large population of speakers to form a general reference model [3],[8]. The probability of the observed test vectors for this general (speaker independent) model (GM) is then used for the normalisation of the likelihood for the target model. It is thought that the effectiveness of this method for reducing false rejection is maximised when speakers are represented using relatively clean reference models. This is because, in this case, the contamination of the test utterance can be expected to give rise to similar levels of mismatch between the test utterance and each of the target and general models. For the purpose of reducing false acceptance, the approach relies on the competitiveness of the adopted general model. However, unlike the competing models used in the other two methods, the employed general (speaker independent) model cannot be expected to be highly similar to either the target model or the test utterance. As a result, it is thought that in terms of false acceptance the approach should be superior to the cohort method but not as effective as the unconstrained cohort technique.

3. SPEECH DATABASE AND ANALYSIS

The speech data used in the experimental study consists of two subsets of the Brent database [9]. Each subset contains repetitions of isolated digit utterances zero to nine. These were collected from telephone calls made from various locations by both male and female English speakers. The first subset consists of 47 repetitions of the above digit utterances spoken by 11 male and 9 female speakers. For each speaker, the first 3 utterance repetitions (recorded in a single call) form the training set. The remaining 44 repetitions (1 recorded per week) are used for testing. The second subset consists of 44 repetitions of the same utterances spoken by another 20 speakers. This subset is used as the speech data from impostors who are outside the set of registered speakers. The general models of the digit utterances are based on the repetitions of these spoken by 100 talkers.

The utterances, which have a sample rate of 8 kHz and a bandwidth of 3.1 kHz, are pre-emphasised using a first order digital filter. These are segmented using a 32 ms Hamming window shifted every 16 ms. Each frame is then appropriately analysed using an 8^{th}-order fast Fourier transform, a filter bank, and a discrete cosine transform to extract an 8^{th}-order mel-frequency cepstral feature vector [10]. The filter bank used for this purpose consists of 19 filters. The centre frequencies of the first 10 filters are linearly spaced up to 1 kHz, and the other 9 are logarithmically spaced over the remaining frequency range (up to 4 kHz). In order to minimise the performance degradation due to the linear filtering effect of the telephone channel, a cepstral mean normalisation approach is adopted. The technique involves computing the average cepstral feature vector across the whole utterance, and then subtracting this from individual feature vectors [9].

4. EXPERIMENTAL INVESTIGATIONS

For the purpose of this study a hidden Markov model (HMM)-based text-dependent speaker verification system is adopted. In this system, speakers are modelled by a set of four-state left to right HMM's representing individual digit utterances. The observation probability for each state is a continuous density function described by a mixture of two Gaussian densities. The covariance matrix of the probability distribution is assumed to be diagonal, and the model parameters are estimated using a modified K-means algorithm [11].

The first part of the experiments is concerned with the evaluation of different score normalisation methods when the impostors are drawn from within the set of registered speakers. In this study the verification scores are expressed in terms of log likelihoods, and the normalised scores are obtained as

$$S_i = S_i^u - S', \tag{3}$$

where

$$S_i^u = \log p(\mathbf{O}|\lambda_i) \tag{4}$$

is the unnormalised verification score for the target speaker *i*. Depending on whether the score

normalisation is based on the use of a cohort of competing speaker models or a general model, S' is given as

$$S' = \frac{1}{N}\sum_{j=1}^{N} \log p(\mathbf{O}|\lambda^{i}_{j}), \quad (5)$$

or

$$S' = \log p(\mathbf{O}|\lambda_{GM}) \quad (6)$$

respectively, where λ^{i}_{j} are the speaker models selected to compete with the model of the target speaker *i*, *N* is the number of these competing models (cohort size), and λ_{GM} is the adopted general model.

In the case of cohort and unconstrained cohort methods, verification trials are performed by first allowing the target model to be included in the set of competing speaker models and then disallowing this. In the former condition, the experiments are conducted by incrementing the cohort size from 1 to 20, and in the latter by incrementing this size from 1 to 19. For the purpose of the cohort method, the selection of the competing models is carried out using the pair-wise comparison technique [2]. In the case of the unconstrained cohort method, as stated earlier, the cohort of competing models is formed during each test trial.

Figure 1 illustrates the results of this experimental comparison in terms of the average equal error rate (EER) for single digit utterances. The EER obtained using unnormalised verification scores is also presented in this figure as the baseline. It is observed that the unconstrained cohort method is considerably more effective than the other two types of score normalisation methods. The superior performance of the unconstrained cohort approach over the cohort technique is particularly significant for small cohort sizes. This is thought to be due to the excellent ability of the unconstrained cohort method to reduce the impostors scores. As the cohort size is increased, the effectiveness of the cohort method improves almost exponentially and the gap between this and the performance of the unconstrained cohort method decreases. Figure 1 shows that for the maximum cohort size, the EERs obtained using these two methods are identical. This is because, in this case, exactly the same competing speakers are used by the two methods.

The results in Figure 1 also indicate that by disallowing the inclusion of the target model in the set of competing models, the performance of the cohort method improves considerably. This improvement appear to be more significant for small cohort sizes. In the case of the unconstrained cohort method, however, the effect is not as noticeable and is almost negligible for cohort sizes of larger than 1.

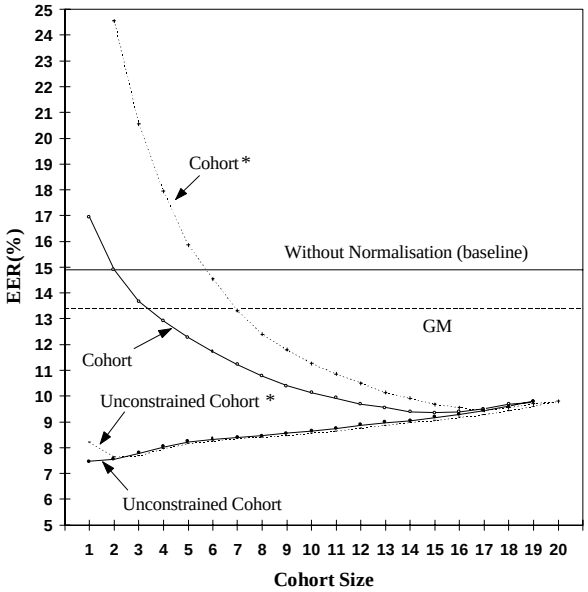

Figure 1. Comparison of various normalisation methods based on the average equal error rate (EER) for single digit utterances.
* Inclusion of the target model in the set of competing models is allowed.

Another interesting aspect of the results in Figure 1 is that, for very small cohort sizes, the EERs obtained using the cohort method are larger than that achieved without normalising the verification scores. These results clearly show that in order for the cohort method to improve the verification accuracy, and also perform better than the GM-based approach, an appropriately large cohort size must be adopted.

4.1. Impostors from Outside the Set

In practical applications of automatic speaker verification the impostors are more likely to be from outside the set of registered speakers. In order to investigate this case, the considered score normalisation methods are used in a set of experiments based on drawing impostors from the second adopted subset of the Brent database. Due to the results obtained earlier, in the case of cohort and unconstrained cohort methods, the experiments are on the basis of excluding the target model from the set of competing models. The results of this study (Figure 2) show that the relative performance of different methods has almost the same pattern as in the previous case. It is observed that, due to its superior ability in reducing verification scores for the impostors, the unconstrained cohort method is again more effective than the other two methods. The approach achieves this ability by allowing the speaker models in the set which are most close to the test utterance to compete with the target model. As a result, provided the set of registered speakers is adequately large, there is always a high

probability that an impostor targeting a particular speaker will score higher against one or more models in the set other than the model of the target speaker.

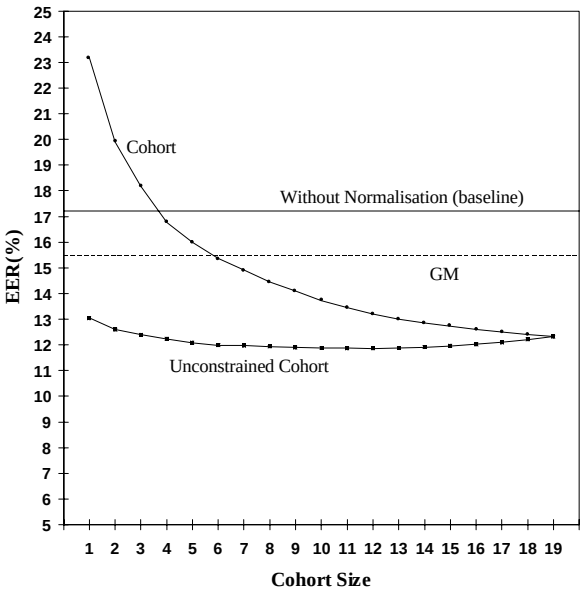

Figure 2. Average equal error rates in experiments using impostors from outside the set of registered speakers.

It should be pointed out that in terms of computational cost, the GM-based method is more efficient than the other two approaches. In the cohort method, The amount of computation involved in calculating the score normalisation term increases linearly with the cohort size. In the case of the unconstrained cohort method, the amount of this computation is linearly related to the size of the set of speakers from which the competing speakers are selected (e.g. the set of registered speakers in this study).

5. CONCLUSIONS

The relative effectiveness of different score normalisation methods for robust text-dependent speaker verification has been experimentally investigated. The study has been based on drawing impostors from within as well as without the set of registered speakers. The experimental results have indicated that, in both cases, the unconstrained cohort method is more effective than either the cohort technique or the approach based on the use of a general model of the utterance. The superior performance of the unconstrained cohort method is due to the fact that the approach allows the speaker models within the set which are most similar to the test utterance to compete with the target model.

The experimental results have also shown that disallowing the inclusion of the target model in the set of competing speaker models considerably improves the effectiveness of the cohort method. In the case of the unconstrained cohort method the effect does not appear to be as significant.

It has been experimentally demonstrated that the performance of the cohort method depends highly on the size of the adopted set of competing models. If this size is too small, then the use of the cohort method in fact leads to less accuracy in speaker verification than that achievable without the normalisation of verification scores. However, by using an appropriately large set of competing models the performance of the method improves significantly, and it even becomes considerably more effective than the general model-based approach.

6. REFERENCES

[1] K-P. Li and J.E. Porter, "*Normalizations and Selection of Speech Segments for Speaker Recognition Scoring*", Proc. ICASSP, pp. 595-598, 1988.

[2] A. E. Rosenberg, J. Delong, C. H. Huang and F. K. Soong, "*The Use of Cohort Normalized Scores for Speaker Verification*", Proc. ICSLP, pp. 599-602, 1992.

[3] M. J. Carey and E. S. Parris, "*Speaker Verification* ", Proc. IOA, Vol. 18, pp. 99-106, 1996.

[4] T. Matsui and S. Furui, "*Concatenated Phoneme Models for Text-Variable Speaker Recognition*", Proc. ICASSP, pp. 391-394, 1993.

[5] N. S. Jayant, "A Study of Statistical Pattern Verification", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. SMC-2, pp. 238-246, Apr. 1972.

[6] S. Furui, "*An Overview of Speaker Recognition Technology*", Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[7] A. M. Ariyaeeinia, P. Sivakumaran and B. Jefferies, "*Speaker Verification in Telephony*", Proc. IOA, Vol. 18, pp. 399-408, 1996.

[8] M. J. Carey and E. S. Parris, "*Speaker Verification Using Connected Words*", Proc. IOA, Vol. 14, pp. 95-100, 1992.

[9] M. Pawlewski, B. P. Milner, S. A. Hovell, D. G. Ollason, S. P. A. Ringland, K. J. Power, S. N. Downey and J. Bridges, "Advances in Telephony Based Speech Recognition", *BT Tech. J.*, pp. 127-150, Jan. 1996.

[10] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, pp. 357-366, Aug. 1980.

[11] L. R. Rabiner, J. G. Wilpon and B-H. Juang, "A Segmental K-Means Training Procedure for Connected Word Recognition", *AT&T Tech. J.*, vol. 65, pp. 21-31, May/June 1986.