Proceedings of the Institute of Acoustics

TEXT-DEPENDENT SPEAKER VERIFICATION UNDER NON-UNIFORM MISMATCHED CONDITIONS

A. M. Ariyaeeinia (1), P. Sivakumaran (1), M. Pawlewski (2) and M. J. Loomes (1)

(1) University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK
(2) BT Laboratories, Martlesham Heath, Ipswich, Suffolk, IP5 7RE, UK

1. INTRODUCTION

The practical applications of speaker verification systems would often involve capturing speech signals in uncontrolled environments and also the transmission of these over communication channels [1][2][3]. As a consequence, speech signals received by verification systems would be adversely affected by the transmission channel and background noise. The resultant variations in speech characteristics can lead to a mismatch between the corresponding training and test utterances which in turn may significantly reduce the verification accuracy. In order to tackle this problem, a number of techniques have been developed in recent years which are mainly based on normalising verification scores [4][5][6][7]. The main drawback of the above methods is that they operate on the assumption that the mismatch is uniform across the given utterance. In practice, however, due to time-localised anomalies in speech the mismatch is mainly non-uniform.

In order to improve the speaker verification performance under non-uniform mismatched conditions, a technique is proposed which is based on a segmented multiple model representation of speakers, and involves evaluating the corrective factors for verification scores by estimating the segmental mismatches. It is demonstrated that through an appropriate evaluation of the required weighting factors, the segmental distances are scaled favourably when the claimant is the true speaker, and unfavourably when this is an impostor. The following sections describe the proposed robust speaker verification technique in detail and present an experimental analysis of its performance.

2. PROPOSED APPROACH

The proposed approach is based on using a dynamic time warping (DTW) algorithm [8], and representing each registered speaker using segmented multiple reference models. Each reference model is formed using a single utterance repetition.

The first part of the technique involves evaluating the relative dissimilarities between each segment of a given test utterance and the corresponding segments in the collection of reference models of the proposed speaker. The best individual reference segments are then selected to form a complete reference model. For the purpose of this process, it is desirable that all the templates for a given utterance text should be of the same length [8]. To achieve this, through the use of a linear decimation-interpolation technique in the training phase, the length of each template is made equal to the mean length of all the available templates for the

given text [9]. This process is repeated during verification trials to ensure that for a given utterance text, the test and reference templates have the same duration.

Figure 1 presents, as an example, the distances (obtained using DTW) between three training and one testing repetitions of a digit utterance spoken by the same speaker. An examination of this figure clearly shows that the relative closeness of the reference templates to the test template vary considerably and irregularly across the length of the utterance. This indicates that by partitioning the utterance into shorter segments, it would be possible to select a set of reference segments (from the given models) with the minimum distances from their corresponding test segments. The distance between the test template and a complete reference model formed using the selected segments can be expressed as

$$D = \frac{1}{K}\sum_{k=1}^{K} \min_{l} d_{k,l}, \qquad 1 \le l \le L \tag{1}$$

where $d_{k,l}$ is the distance between the k^{th} segment of the test utterance and the corresponding segment in the l^{th} reference model, K is the number of segments, and L is the number of reference models.

An important issue to consider in this approach is the size of the segments. Based on the graphs in Figure 1, it can be argued that in order to minimise the overall distance, the segments should have the shortest possible length. The use of DTW in fact provides the possibility of reducing the segment size to that covering only a single frame. In this case, the overall distance can be obtained as the average of distances between the test utterance frames and the best corresponding frames in the set of reference models. A trace of these distances is shown in Figure 1. With this size of segments, the approach may also be viewed as a DTW technique with a three-dimensional search for the optimum path.

Figure 1. Plots of distances between three utterance repetitions in the reference set and a test utterance spoken by the same speaker. (i): Trace of minimum distances.

The second part of the proposed approach is concerned with reducing the effects of any existing mismatch between the test utterance and the generated best reference template. This is achieved by weighting each segmental distance in accordance with an estimated level of mismatch associated with that segment. The overall distance is then computed as the average of these weighted segmental distances, i.e.

$$\hat{D} = \frac{1}{N} \sum_{n=1}^{N} w(n) d(n) , \tag{2}$$

where N is the adjusted number of frames in the given utterance, $d(n)$ is the distance between the n^{th} test frame and the corresponding frame in the generated best reference model, and $w(n)$ is the weighting factor for the n^{th} segmental distance. In order to determine the required weighting factors, use can be made of a set of speaker models that are capable of competing with the target model. Such competing speaker models can be selected in a manner similar to that in the cohort normalisation method [7], that is on the basis of their closeness to the target model. In this case, the weighting function can be defined as:

$$w(n) = \left[\frac{1}{J} \sum_{j=1}^{J} d_j^{'}(n) \right]^{-1} , \tag{3}$$

where J is the number of speakers in the selected competing set, and $d_j^{'}(n)$ is the distance between the n^{th} segment of the test utterance and the best corresponding segment in the collection of reference models associated with the j^{th} competing speaker. Unlike the conventional cohort normalisation (CN) method, this approach involves applying a segmental selection procedure to the reference models of each pre-selected competing speaker. As a result, a pre-selected competing speaker may not necessarily be represented by exactly the same model in different verification trials. This method of calculating segmental weights is illustrated in Figure 2 (a). It should be pointed out that this approach has the same disadvantage as the conventional cohort normalisation method [4]. That is, if an impostor produces a test utterance which is almost equally dissimilar from the proposed model and competing models, then the approach may lead to a small overall distance, and hence accept the impostor as the true speaker. This is simply because, in this case, the large segmental distances given by $d(n)$ are almost cancelled out by the small values of $w(n)$.

A method for tackling the above problem is to choose the competing speaker models based on their closeness to the given test template. With this method, when the test utterance is produced by the true speaker, the competing speaker models can be assumed to be adequately close to the true speaker reference model. Therefore the method can be expected to be almost as effective as the previous approach. However, in the case of the test utterance being produced by an impostor, the competing speaker models will be similar to the test template and not necessarily to the target model. As a result $d(n)$ and $w(n)$ will both become large and the possibility of false acceptance will be reduced significantly. From the point of view of

selecting competing speaker models, this approach can be seen as a modified form of the unconstrained cohort normalisation (UCN) method [4] in which competing speaker candidates are represented using the best possible models in terms of closeness to the test utterance. To achieve this, the multiple reference models for each competing speaker candidate are subjected to the same segmental selection procedure as that in the case of the proposed speaker. This method of calculating segmental weights is summarised in Figure 2 (b). For the purpose of this paper the above two methods are referred to as segmental weighting type 1 and segmental weighting type 2 respectively.

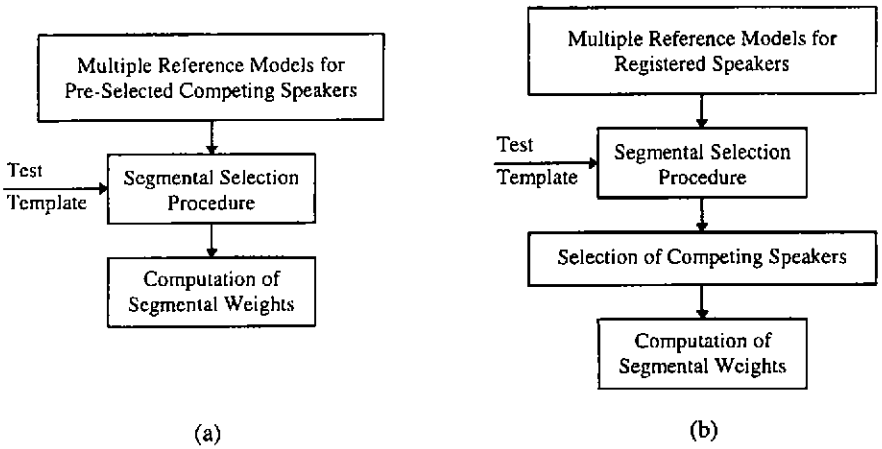

Figure 2. Computation of the required segmental weights; (a) using pre-selected competing speakers, and (b) for reducing both false rejection and false acceptance rates.

3. SPEECH DATA AND ANALYSIS

For the purpose of this study a subset of the Brent speech database consisting of 47 repetitions of isolated digit utterances zero to nine is adopted [3]. This subset was collected from telephone calls made from various locations 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 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 25 ms Hamming window shifted every 12.5 ms, and then subjected to a 12^{th}-order linear prediction analysis.

The resultant linear predictive coding (LPC) parameters for each frame are appropriately analysed using a 10^{th}-order fast Fourier transform, a filter bank, and a discrete cosine transform to extract a 12^{th}-order mel-frequency cepstral feature vector [10]. The filter bank used for this purpose consists of 20 filters. The centre frequencies of the first 10 filters are linearly spaced up to 1 kHz, and the other 10 are logarithmically spaced over the remaining frequency range.

4. EXPERIMENTAL WORK

The first part of the investigation is aimed at determining the relative effectiveness of the methods described in section 2. For this purpose, three sets of experiments are conducted using the proposed segmental selection approach. The first set of these is carried out without the use of any segmental weighting, whereas the second and third sets are based on the use of the segmental weighting type 1 and type 2 methods respectively. The baseline in this study is the speaker verification performance obtained by representing each registered speaker using a combined reference model [8]. The sets of competing speakers used in this and other investigations presented in the paper, are all of size ten.

The results of this experimental study are presented in terms of the average equal error rate (EER) for single digit utterances and the EER for a combination of all ten digits in Figures 3 (a) and (b) respectively. These results clearly confirm the effectiveness of the segmental selection approach for reducing the verification error, particularly when it is combined with one of the proposed segmental weighting schemes. The superior performance of the segmental weighting type 2 over type 1 is, as stated earlier, due to its additional ability to increase the segmental distances when the claimant is an impostor. The two weighting methods, however, are almost equally effective in reducing the adverse effects of mismatch when the claimant is the true speaker.

The results given in Figure 3 are based on averaging all the segmental distances for a given test utterance. It is, however, thought that by excluding the few largest segmental distances from the computation of the overall distance, a higher accuracy in verification may be obtained. In the case of the test utterance being produced by the true speaker, such large segmental distances can be due to a high level of mismatch between corresponding segments of the test and reference templates. To examine the effects of discarding larger segmental distances, a set of experiments is conducted with the proposed methods, and by using single digit utterances as inputs. In these experiments the number of discarded segmental distances is incremented from zero to a maximum of 15. The results of this study (Figure 4) show that when the segmental selection technique is implemented on its own, discarding the few largest segmental distances can lead to some improvement in the verification accuracy. In the case of SSWT1, however, the reduction achieved in the EER through this method appears to be far less significant. This is thought to be mainly due to the effectiveness of the adopted weighting scheme in compensating for any segmental mismatch. It is also observed that, in the case of

the SSWT2 approach, it is not advisable to exclude larger segmental distances from the calculation of the final distance as this leads to an increase in the verification error. This is believed to be due to the additional ability of SSWT2 to increase the segmental distances when the test utterance is produced by an impostor. As a result, discarding larger segmental distances can adversely affect the impostor rejection capability offered by the approach.

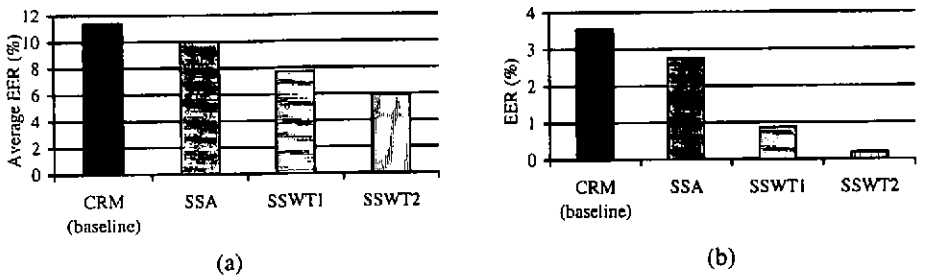

Figure 3. Effectiveness of the proposed methods in terms of (a) the average EER for single digit utterances, and (b) the EER for a combination of all ten digit utterances.
CRM = Combined Reference Model.
SSA = Segmental Selection Approach.
SSWT1 = Segmental Selection and Weighting Type 1.
SSWT2 = Segmental Selection and Weighting Type 2.

In order to obtain a more meaningful evaluation of the performance of the proposed approach, the SSWT1 and SSWT2 methods are experimentally compared with the cohort and unconstrained cohort normalisation techniques respectively. This choice of paring is based on the similarity in selecting competing speakers. Table 1 gives the results of this study in terms of the average EER for single digit utterances and the EER for a combination of all ten digits. It is observed that SSWT1 and SSWT2 have very similar levels of superior performance over CN and UCN respectively. The results also show the SSWT2 method as the best performer. The average EER of 5.92% for single digit utterances and the EER of 0.19% for a combination of all ten digit utterances obtained with this method are found to be particularly encouraging.

The only drawback of the SSWT2 method appears to be its high computational complexity. The reason is that, in this method, a large number of DTW-based template comparisons have to be carried out in order to select the competing speakers. This problem can, to a certain extent, be overcome by selecting the competing speakers through a method which is computationally more efficient. It should, however, be noted that the technique used to replace DTW for selecting competing speakers may not be as efficient. It is therefore possible that the selected competing speakers are different from those that should, and would be obtained with

DTW. An alternative approach would be to use a computationally efficient method to select a larger than required number of competing speakers and then reduce this using DTW.

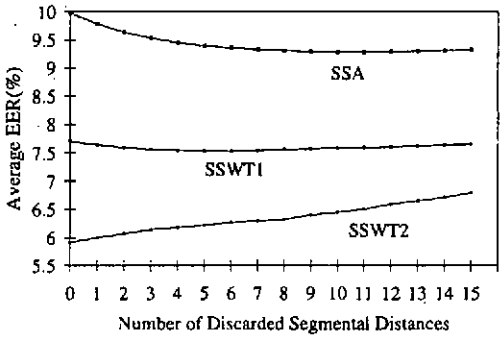

Figure 4. Average EERs for different methods as a function of number of discarded segmental distances.

Method	Ave. EER (%) for Single Digits	EER (%) for a Combination of all ten Digits
CN	8.89	1.13
SSWT1	7.70	0.88
UCN	7.17	0.41
SSWT2	5.92	0.19

Table 1. Comparison of the proposed methods with the conventional cohort normalisation techniques.

5. SUMMARY AND CONCLUSION

An investigation into a method for robust text-dependent speaker verification has been presented. The proposed approach, which is based on a segmented multiple model representation of speakers, consists of two stages. The first stage involves selecting the best segment from the collection of models for the proposed speaker, for each segment of the given test utterance. The selected segments are then used to form a complete reference model for the purpose of verification. The next stage is concerned with weighting the individual segmental distances with the primary aim of compensating for any mismatch between each

corresponding pair of segments in the test and reference templates. As a result, for a fixed verification threshold, the possibility of false rejection is reduced considerably. It has, however, been shown that through the use of an appropriate method for calculating the required weighting factors, the possibility of false acceptance can also be reduced significantly. Based on experiments using telephone quality speech, it has been demonstrated that the proposed approach is considerably more effective than the conventional methods for text-dependent speaker verification under mismatched conditions.

6. REFERENCES

[1] Ariyaeeinia A.M., Sivakumaran P. and Jefferies B., "Speaker Verification in Telephony", *Proc. IOA* (UK), vol. 18, pp. 399-408, 1996.

[2] Naik J.M., Netsch L.P., and Doddington G.R., "Speaker Verification Over Long Distance Telephone Lines", *Proc. ICASSP*, pp. 524-527, 1989.

[3] Pawlewski M. and Downey S., "Channel Effects in Speaker Recognition", *Proc. IOA* (UK), vol. 18, pp. 115-122, 1996.

[4] Ariyaeeinia A.M. and Sivakumaran P., "Analysis and Comparison of Score Normalisation Methods for Text-Dependent Speaker Verification", *Proc. Eurospeech*, vol. 3, pp. 1379-1382, 1997.

[5] Carey M.J. and Parris E.S., "Speaker Verification", *Proc. IOA* (UK), vol. 18, pp. 99-106, 1996.

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

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

[8] Ariyaeeinia A.M. and Sivakumaran P., "Comparison of VQ and DTW Classifiers for Speaker Verification" *Proc. IEE-ECOS* (UK), pp. 142-146, 1997.

[9] Myers C.S., Rabinar L.R., and Rosenberg A.E., "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition", *IEEE Trans. ASSP*, vol. 28, pp. 622-733, 1980.

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