Proceedings of the Institute of Acoustics

AN EFFECTIVE SUB-BAND BASED APPROACH FOR ROBUST SPEAKER VERIFICATION

P. Sivakumaran, A. M. Ariyaeeinia, J. A. Hewitt and J. A. Malcolm

University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK

1. INTRODUCTION

The concept of splitting the entire frequency domain into sub-bands and processing these independently in between every consecutive recombination stage to generate a global score has already been investigated for speech recognition [1][2]. Some aspects of this technique have also been studied for the task of speaker recognition [3][4].

The main motivation for the above approach is that it allows for selective de-emphasis of sub-bands that are affected by narrow band noise and it permits emphasis of the sub-bands which are more specific to the speaker. It also provides the possibility of relaxing the conventional time-synchrony assumption between the sub-bands [1][5]. Moreover, the approach allows a closer simulation of the human perception [6].

The main issue addressed in this paper is the reduction of the effects of any existing mismatch between the band-limited segments of the test and reference utterances in a sub-band based speaker verification system. This can be achieved by using a weighting scheme which ensures that the scores associated with corrupted band-limited segments are appropriately de-emphasised. The weighting factors required for this purpose can be computed using segmental scores obtained for a set of background speaker models. The general idea behind this approach is that if due to certain time and frequency localised anomalies there is some degree of mismatch between a particular band-limited segment of the test utterance (produced by the true speaker) and the corresponding segment of the target model, then a similar level of mismatch should exist between the considered test segment and the corresponding segments of the background speaker models.

It is believed that through an appropriate selection of background speaker models, the above weighting scheme may lead to the emphasis of the sub-bands that are more specific to the target speaker. The idea is based on the view that the mean separation between the scores of the target and background speaker models for a particular sub-band is a measure of the performance of that sub-band for the given target speaker.

The paper also includes a study of two other aspects of the sub-band approach which have not been investigated for speaker recognition previously. These are the relaxation of the conventional time synchrony assumption of different sub-bands and the difficulties associated with the sub-band cepstral features.

This paper is organised in the following manner. The next section details the classification process used in this work and describes the adopted merging strategy. Section 3 gives a description of the utilised speech database, and the method used for the extraction of sub-band feature vectors. The experimental work and results are detailed in Section 4, and the overall conclusions are presented in Section 5.

2. CLASSIFICATION AND MERGING

The technique used for this purpose can be thought of as a set of simultaneous dynamic time warping (DTW) processes in which the distance accumulation involved in every individual process is affected by the others at certain intermediate levels. Each DTW process is, in fact, associated with a different sub-band and the intermediate effects are due to the sub-band recombination process. Figure 1 represents the concept involved in this approach in a three dimensional domain.

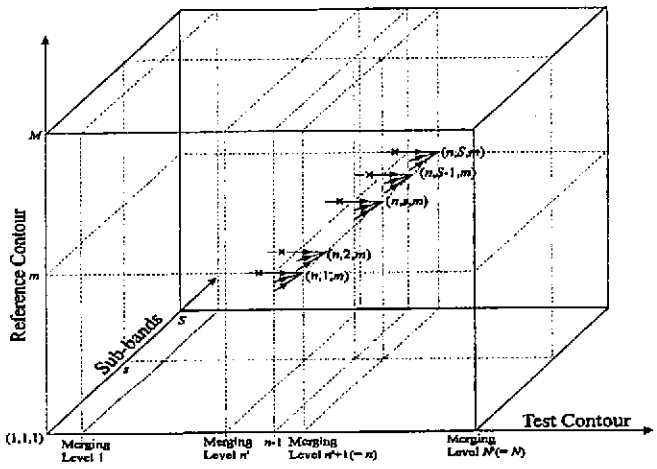

Figure 1 : Utilised classification and merging technique.

In the above illustration, it is assumed that the test and reference contours are represented as $T(n,s)$, $1 \le n \le N$, $1 \le s \le S$ and $R(m,s)$, $1 \le m \le M$ respectively, where N and M denote the lengths of the corresponding utterances, and S represents the number of sub-bands.

In order to formulate the above problem, the relaxed endpoint and the Itakura's local path constraints are adopted [7]. It should be pointed out that these constraints will impose an additional global path restriction on each n-m plane. With these constraints a three step procedure can be derived to obtain the required solution.

Step 1 : Initialisation :

From $m = 1$ to $1+\delta$ and for all s

$$\begin{aligned} D_A(1,s,m) &= \frac{1}{S}\sum_{i=1}^{S} d(1,i,m)h(1,i) \quad \text{if } n = 1 \text{ is a merging level} \\ &= d(1,s,m) \qquad\qquad\qquad \text{otherwise} \end{aligned} \tag{1}$$

Step 2 : Main Recursion :

From $n = 2$ to N, $m = 1$ to M and for all s

If $M_L(n) \le m \le M_H(n)$ then

$$D_A(n,s,m) = d(n,s,m)h(n,s) + \min\begin{Bmatrix} D_A(n-1,s,m)g(n-1,s,m), \\ D_A(n-1,s,m-1), \\ D_A(n-1,s,m-2) \end{Bmatrix} \quad (2)$$

If n is a merging level :

$$D_A(n,s,m) = \frac{1}{S}\sum_{s=1}^{S} D_A(n,s,m) \quad (3)$$

Step 3 : Termination : (Final score)

$$D = \min_{M-\delta \le M_e \le M}\left[\frac{1}{SN}\sum_{s=1}^{S} D_A(N,s,M_e)\right] \quad (4)$$

In the above procedure δ is the maximum anticipated range of mismatch (in frames) between boundary points of the considered utterances, $h(n,s)$ is a weighting factor which is associated with the n^{th} test frame of the s^{th} sub-band, $d(n,s,m)$ is a weighted Euclidean distance between the n^{th} test frame of the s^{th} sub-band and m^{th} reference frame of the corresponding sub-band, $M_L(n)$ and $M_H(n)$ are the lower and upper boundaries of the global constraint respectively and have the forms $M_L(n) = \max[0.5(n+1), M-2(N-n)-\delta, 1]$, $M_H(n) = \min[2n+\delta-1, M-0.5(N-n), M]$, and $g(\cdot)$ is the conventional non-linear scaling factor which prevents the optimum path to be flat for two consecutive frames [7]. In order to make the above procedure equally effective for all ratios of N/M, a linear decimation-interpolation technique is adopted to make the length of the test vector sequences equal to that of the reference [8]. For the purpose of this paper the above approach is referred to as simultaneous DTW (SDTW).

The use of weighting factors $h(n,s)$, $1 \le n \le N$, $1 \le s \le S$ as described in the above formulation, provides the possibility of correcting each band-limited segmental distance in accordance with the associated level of mismatch. In order to determine these weighting factors, use can be made of either the speaker independent sub-band models or a set of sub-band speaker models that are capable of competing with the target model. In the latter case the required competing speaker models can be selected based on their closeness to either the target model or the test utterance [9]. For the reason stated below, the second approach was chosen for this work. Based on this technique, an effective weighting function can be defined as

$$h(n,s) = \left[\frac{1}{J}\sum_{j=1}^{J} d_j(n,s)\right]^{-1}, \quad (5)$$

where J is the number of speakers in the selected competing set, $d_j(n,s)$ is the distance between the n^{th} test frame of the s^{th} sub-band and optimally aligned frame of the j^{th} competing speaker model. The above formulation implies that the SDTW and a backtrack procedure have to be applied for each combination of the test template and the competing speaker model. It should be noted that each of these procedures involves a

different set of $h(\cdot)$. These weighting factors can simply be specified as a common constant, or computed according to signal-to-noise ratios (SNRs) in the band-limited frames. For the latter approach, an estimation of band-limited noise can be obtained either during the silence periods or by using spectral magnitude distributions of the band-limited speech segments (in this case relatively large speech segments are required) [10]. It should be pointed out that the SNR based weighting factors may be effective only when the distortion is due to relatively stationary additive noise. This implies that the technique may not be useful for minimising the effects of such causes of mismatch as the speaker generated variation or convolutional noise.

The main attraction of the adopted approach for choosing the competing speaker models is its excellent ability to reduce the false acceptance error [9]. This is because when the test utterance is 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 $d(n,s)$ and $h(n,s)$ both will become large and thereby the probability of false acceptance will be reduced significantly. Figure 2 illustrates the main operations involved in the proposed approach.

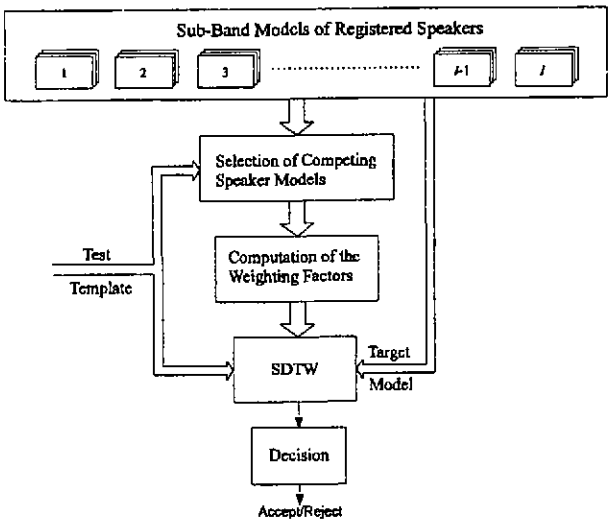

Figure 2 : Proposed verification method.

In order to perform a meaningful evaluation of the effectiveness of the proposed method, it should be compared with a full-band approach which is expected to be robust against adverse effects. One such technique is that based on the standard DTW and unconstrained cohort normalisation (UCN) [9]. A fundamental difference between the proposed SDTW approach and the above full-band (DTW+UCN) method is that the latter assumes that the mismatch is uniform across the given utterance. The SDTW technique, on the other hand, does not make such an assumption and attempts to estimate the level of mismatch associated with individual band-limited utterance segments. This information is then used to compute a weighting factor for correcting each segmental distance prior to the calculation of the final distance.

3. SPEECH DATA AND FEATURES

The speech data used for this study was a subset of the BT Millar speech database. The subset consisted of 25 repetitions of digit utterances one to nine and zero spoken by 20 male speakers of about the same age. The first 10 versions of each utterance were reserved for training and the remaining 15 formed the standard test set. The adopted subset was recorded in a quiet environment and, had a bandwidth of 3.1 kHz and a sample rate of 8.0 kHz.

In the experimental study two different sets of sub-band features were considered. These were SB-MFBOs and SB-MFCCs (here the abbreviations SB, MFBOs and MFCCs stand for sub-band, mel-scale filterbank outputs and mel frequency cepstral coefficients respectively). In order to generate these features, the utterances were first pre-emphasised using a first-order digital filter. Each utterance was then segmented into 32 ms frames at intervals of 16 ms using a Hamming window, and subjected to an 8^{th} order fast Fourier transform (FFT). The resulting energy spectrum for each frame was analysed appropriately using a mel-scale filterbank [11]. The frequency range was divided into four overlapping sub-bands covering the frequency intervals 0-600 Hz, 500-1149 Hz, 1000-2297 Hz, and 2000-4000 Hz. The log-energy outputs of the filterbank were then grouped according to these sub-bands to obtain SB-MFBOs. In order to compute SB-MFCCs a discrete cosine transform (DCT) was applied to each group of SB-MFBOs.

The full-band feature sets which were used for the purpose of comparative studies were MFBOs and MFCCs. The former was a cascade of the corresponding groups of SB-MFBOs and the latter was obtained by applying a DCT to the resultant set of MFBOs.

4. EXPERIMENTAL WORK AND RESULTS

The first set of experiments was conducted to determine the best possible level to perform the sub-band merging process. For this purpose, SB-MFBO features were used. The merging levels considered were frame, phoneme and word. In order to obtain the required phonetic boundaries in the second case, the reference templates were forced aligned against phoneme-based hidden Markov models (HMM) of the corresponding utterance text. Table 1 presents the results of this study in terms of equal error rate (EER). Although these results are in favour of the frame-level merging, it should be noted that the difference in the performance between the frame and phoneme-level merging is not significant. An additional set of experiments has shown that changing the merging level from a single frame to about 3-5 frames (depending on the utterance text), leads to an improvement of around 10-15% in the EER. This result was unexpected since the recombination is expected to be meaningful at a time-resynchrony point [1][5]. Thus, further investigations should be carried out in order to make any conclusion. In the remaining part of the experimental work the single frame level merging was used.

Merging level	Single-frame	Phoneme	Word
EER (%)	14.11	14.67	15.78

Table 1 : EER for different merging levels.

The next set of experiments was conducted using the SB-MFBO and MFBO features to compare the effectiveness of the proposed SDTW technique against that of the full-band approaches (i.e. standard DTW and DTW+UCN). In this study, an adverse effect was simulated by contaminating 1/3 of the test utterances with a narrow band noise (0-600 Hz). The results of this investigation are presented as a function of SNR in Figure 3. These results clearly confirm the robustness of the SDTW method in the considered condition. The figure also shows the benefit of introducing the score normalisation in the full-band approach.

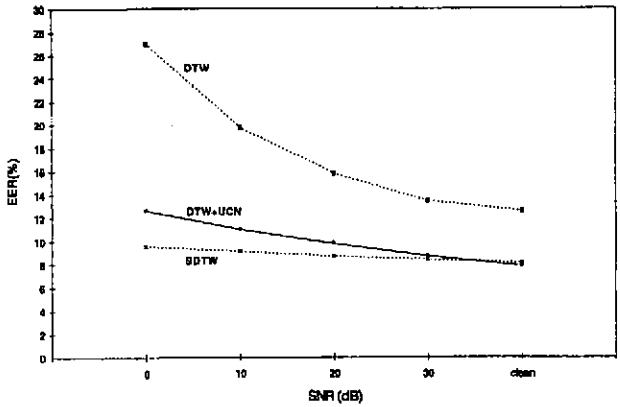

Figure 3 : Performance of the considered full and sub-band approaches as a function of SNR. The feature parameters are MFBOs for the full-band methods, and SB-MFBOs for the sub-band approach.

The above experiments were repeated using the SB-MFCC and MFCC features. The results of this investigation are given in Figure 4. As before, the SDTW method exhibits a relatively flat response across the considered SNR range. However, the overall performance of the DTW+UCN is noticeably better than that of the SDTW approach. This may be due to the way SB-MFCCs are generated. As described earlier, the generation of these parameters involves the use of independent DCTs in each sub-band. The purpose of this is to obtain a separate set of uncorrelated features for individual sub-bands. However, this can also result in a more detailed representation of the overall spectral envelope variations [12]. For example, in the case of four sub-bands, the details of the overall spectral variations measured by the 1st and 2nd DCT basis functions of individual sub-bands are, to a large extent, similar to those of 4th and 8th DCT basis functions of the full-band respectively (Figure 5). This implies that an alternative subset of SB-MFCCs does not exist to represent details of the spectral variations that are described by any of the full-band MFCCs 1-3, 5-7, and so forth.

In order to deal with the above problem a modified SDTW (MSDTW) method is considered. This approach involves the use of a set of full-band MFCCs that are not represented by SB-MFCCs, i.e. a form of complementary features to SB-MFCCs (hence they are referred to as CMFCCs). In order to use these features in the SDTW procedure, the term $\sum_i d(n,s,m)h(n,s)$ in equation (1) and also in the subsequent part of the algorithm (when the merging is carried out at the frame level) is replaced by

$$\alpha \sum_{s=1}^{S} d(n,s,m)h(n,s) + (\alpha - 1)d'(n,m)h'(n)$$

where α is a combination factor between 0 and 1, $d'(n,m)$ is a weighted Euclidean distance between the CMFCCs of the test and reference utterances, and $h'(n)$ is a weighting factor which is computed using the CMFCCs of the competing speaker models. The use of these weights provides the possibility of correcting each segmental distance in accordance with the associated level of mismatch. It is clear that, due to the involvement of the full-band features, the benefits of the sub-band processing cannot be fully realised. However, the experimental results (with $\alpha = 0.5$) presented in Figure 4 imply that significant improvement in EER can be achieved by using these full-band features in the manner described.

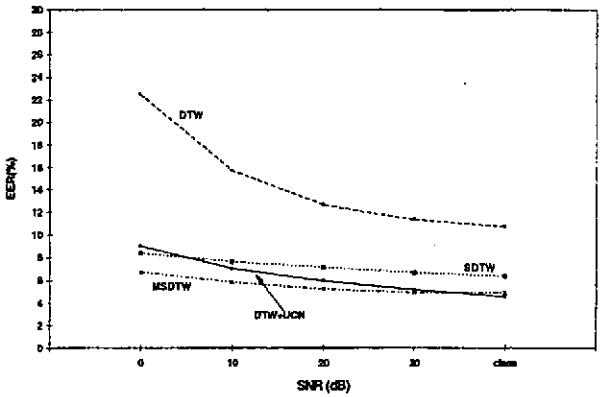

Figure 4 : Performance of the considered full and sub-band approaches as a function of SNR using the adopted cepstral features.

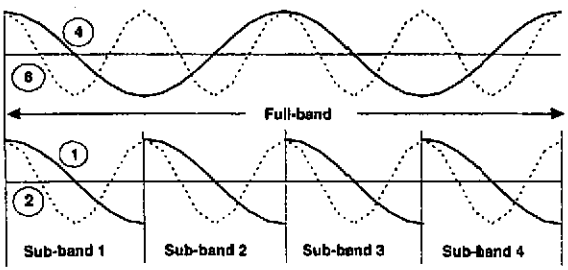

Figure 5 : Comparison between the full and sub-band DCT basis functions.

5. CONCLUSION

A sub-band technique for robust text-dependent speaker verification has been investigated. The proposed approach attempts to reduce the effects of mismatch between the band-limited segments of the test and reference material by using an appropriate weighting scheme. The weighting factors required for this purpose are obtained using a set of competing speaker models. The effectiveness of the approach was clearly observed in the experimental study conducted using SB-MFBOs. However, this result was not repeated when SB-MFCCs were used. The reason for this was found to be the lack of spectral information in SB-MFCCs. This difficulty was, to a cretin extent, overcome by using a set of complementary features. Finally, it should be pointed out that although the experimental work was carried out using narrow band noise, the proposed approach is capable of handling any form of undesired mismatches which are due to the time or/and frequency-localised anomalies.

6. REFERENCES

[1] H. Bourlard and S. Dupont, "A new ASR approach on independent processing and recombination of partial frequency bands," *Proc. ICSLP'96*, pp. 426-429, 1996.

[2] H. Hermansky, S. Tibrewala and M. Pavel, "Towards ASR on partially corrupted speech," *Proc. ICSLP'96*, pp. 462-465, vol. 1, Oct. 1996.

[3] L. Besacier and J. Bonastre, "Subband approach for automatic speaker recognition : optimal division of the frequency domain," *Proc. AVBPA'97*, pp. 195-202, 1997.

[4] R. Auckenthaler and J. S. Mason, "Equalizing sub-band error rates in speaker recognition," *Proc. Eurospeech'97*, pp. 2303-2306.

[5] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland and M. A. Fawley, "Modelling asynchrony in speech using elementary single-signal decomposition," *Proc. ICASP'97*, pp. 1247-1250, 1997.

[6] J. B. Allen, "How do human process and recognize speech ?," *IEEE Trans. on speech and audio processing*, vol. 2, no. 4, pp. 567-577, 1994.

[7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on ASSP*, vol. 29, pp. 254-272, April 1981.

[8] C. S. Myers , L. R. Rabinar, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. on ASSP*, vol. ASSP-28, pp. 622-733, Dec. 1980.

[9] A. M. Ariyaeeinia and P. Sivakumaran "Analysis and comparison of score normalisation methods for text-dependent speaker verification," *Proc. of Eurospeech'97*, pp. 1379-1382.

[10] H. G. Hirsch, "Estimation of noise spectrum and its applications to SNR estimation and speech enhancement," *Tec. Rep. TR-93-012*, ICSI, Berkeley CA. 1993.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, pp. 357-366, Aug. 1980.

[12] S. Vaseghi, N. Harte and B. Milner, "Multi-resolution phonetic/segmental features and models for HMM-based speech recognition," *Proc. ICASP'97*, pp. 1263-1266, 1997.