SUB-BAND BASED SPEAKER VERIFICATION USING DYNAMIC RECOMBINATION WEIGHTS

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

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

ABSTRACT

This paper describes a new method for generating the recombination weights in sub-band based speaker verification. The approach, which is based on the use of background speaker models, attempts to reduce the effect of any mismatch between the band-limited segments of the test utterance and the corresponding sections in the target speaker model. The discussion also includes an analysis of other possible methods for determining these weights. Moreover, a problem generally associated with the sub-band cepstral features is pointed out and a possible solution is presented.

1. INTRODUCTION

The concept of splitting the entire frequency domain into sub-bands and processing the spectra in these bands independently in between every consecutive recombination stage to generate a final score has recently been proposed for speech recognition [5][7]. Some of the aspects of this technique have also been studied for the task of speaker recognition [3][4].

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

A critical issue in the sub-band based approach is the determination of recombination weights. This paper introduces a novel method for generating these weights for the purpose of speaker verification. The technique is based on the use of a set of background speaker models. The underlying idea is to obtain a set of dynamic or run-time recombination weights for each sub-band based on the argument 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 section in the target model, then a similar level of mismatch should exist between the considered test segment and the corresponding sections in the background speaker models.

It is believed that through an appropriate selection of the 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 is organised in the following manner. The next section presents the sub-band based speaker verification method used in this work. Section 3 describes conventional techniques for estimating the recombination weights and then details the proposed method. Section 4 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 5, and the overall conclusions are presented in Section 6.

2. ADOPTED APPROACH

For the purpose of this study, each registered speaker is represented using a set of hidden Markov models (HMMs) in which each model is formed separately in different sub-bands using the standard training algorithm. Moreover, the Viterbi algorithm is modified as follows (this modified version is referred to as sub-band Viterbi, i.e. SBV, algorithm) :

Step 1 : Initialisation :

$$\delta_1(1) = \frac{1}{S}\sum_{s=1}^{S} w_1(s)\log b_{s1}(O_{s1}) \quad (1)$$

$$\text{for } j = 2 \text{ to } N, \delta_1(j) = -\infty \quad (2)$$

Step 2 : Main Recursion : for $t = 2$ to T and $j = 1$ to N

$$\delta_t(j) = \frac{1}{S}\sum_{s=1}^{S}\left\{\max_{1\le i\le N}\left[\delta_{t-1}(i) + \log a_{sij}\right] + \log\left(w_t(s) b_{sj}(O_{st})\right)\right\} \quad (3)$$

Step 3 : Termination : final score

$$l = \max_{1\le j\le N}\left[\delta_T(j)\right] \quad (4)$$

where a_{sij} are the state transition probabilities associated with the s^{th} sub-band model, $b_{sj}(O_{st})$ is the probability for observing the t^{th} test vector of the s^{th} sub-band in the j^{th} state of the s^{th} sub-band model, N is the number of states in each sub-band model, T is the number of test vectors in each sub-band, S is the number of sub-bands and $w_t(s)$ are the recombination weights which will be defined in the next section. It should be pointed out that in the above formulation it is inherently assumed that the sub-band recombination is at the frame level. This is because a set of preliminary experiments has indicated that such a recombination level yields the highest recognition accuracy in speaker verification (a similar result has been reported for speech recognition [7]).

3. ESTIMATION OF RECOMBINATION WEIGHTS

This section focuses on the possible methods for estimating the required recombination weights. The discussions start with the adaptation of certain existing techniques for this purpose [5][7] [9]. A novel approach is then introduced which is shown to be considerably more effective.

3.1. Use of a Priori Knowledge of the Sub-Band Performance

A method for computing the required weights is based on the knowledge of the relative performance of sub-bands which is gained through a series of experiments using a given set of speech data. For example, if the average verification rates for the sub-bands 1,2,...,S are $r_1,r_2,...,r_S$ respectively at a given speech unit level (e.g. phoneme, syllable, word), then the required weights may be specified as :

$$w_t(s) = r_s \Big/ \sum_{s=1}^{S} r_s \qquad \begin{matrix} 1 \le s \le S \\ T' \le t \le T'' \end{matrix} \tag{5}$$

where T' and T'' are the boundaries of the considered speech unit. The above formulation implies that the weights are linearly proportional to normalised verification rates. Alternatively, based on the argument that such linear schemes may not be the most effective approach for this purpose, the verification rates may be used in a non-linear procedure to compute the required weights. For example

$$w_t(s) = \mathrm{f}\left(r_s / \max(r_s)\right) \Big/ \sum_{s=1}^{S} \mathrm{f}\left(r_s / \max(r_s)\right) \qquad \begin{matrix} 1 \le s \le S \\ T' \le t \le T'' \end{matrix} \tag{6}$$

where $\mathrm{f}(x) = x / \left(1 + 20e^{-6x}\right)$. (7)

Comparing this approach with that described by equation (5), it can be seen that in this case, the sub-bands with higher relative verification rates are weighted more heavily. Another, perhaps more effective, mechanism for the non-linear recombination of sub-bands is the discriminative training method [5] [7]. In this technique the required weights are chosen in such a way that the rate of misclassifications is minimised for the given set of data.

In general, the above approaches are expected to improve the verification accuracy by appropriately emphasising the sub-bands that are more specific to the target speaker. However, since the weights are computed prior to the verification process, if a test utterance (produced by the true speaker) is contaminated in the regions where the weights are relatively high, then the techniques can lead to an increase in the false rejection error. An obvious way of tackling this problem is to incorporate the contamination level of the test utterance into the process of generating the weights. The techniques of this category are described in the following sections.

3.2. Use of the Segmental SNR in Each Sub-Band

In order to reduce the effect of additive, band-limited noise, the recombination weights may be computed as SNR dependent. An important issue in this approach is the estimation of the noise levels. A common method for this purpose is the use of the noise spectrum in the last few non-speech segments preceding the speech utterance. In such an approach, the required weights can be specified as follows :

$$w_t(s) = \frac{1}{S-1} \left\{ 1 - \left(\Phi(t,s) \Big/ \sum_{s=1}^{S} \Phi(t,s) \right) \right\} \tag{8}$$

where $\Phi(\cdot)$ is a non-linear function which controls the heaviness of weights according to the local SNR. A number of possible types of this function can be derived from the theory of spectral subtraction [9]. An example of this is

$$\Phi(t,s) = \frac{1}{K'' - K' + 1} \sum_{k=K'}^{K''} \frac{B_{\max}(k)}{1 + \gamma\rho(t,k)} \tag{9}$$

where K'' and K' are the indices of the upper and lower frequency boundaries of the s^{th} sub-band, γ is a scaling factor, $B_{\max}(k)$ is the maximum noise magnitude at k^{th} frequency index in the considered noise frames, and the $\rho(\cdot)$ is the presumed band-limited frame level SNR which is given as :

$$\rho(t,k) = |X(t,k)| / |B(t,k)| \tag{10}$$

where $|X(\cdot)|$ and $|B(\cdot)|$ are the estimates for the spectral magnitudes of the smoothed noisy speech and the noise respectively [9].

The main assumption in the above approach is that the interfering noise remains stationary during speech activities. This, however, cannot be the case in many practical applications. In order to tackle the problem, an approach has been proposed in [8]. The technique involves the use of spectral magnitude distributions of the band-limited speech segments. The estimation of the noise levels is in fact based on the peak shifts observed in these distributions. A disadvantage in this method is that, for accurate estimation of the noise level, a relatively large speech segment (typically in the range of 0.5-2.0 s) is required. The technique proposed in the present work not only deals with this problem effectively, but also handles the effects of various other forms of undesired mismatches that may be speaker generated or due to the environmental and communication channel noise.

3.3. Dynamic Recombination Weights (DRW)

As described earlier, in order to determine the recombination weights according to the level of mismatch between the band-limited segments of the test utterance and corresponding sections in the sub-band models of the target speaker, 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 [2]. For the reason stated below, the second approach was chosen in this study. Based on this method, recombination weights can be defined as

$$\log w_t(s) = -\frac{1}{M} \sum_{m=1}^{M} \log b^m_{q(s,t)}(O_{st}) \tag{11}$$

where M is the number of speakers in the selected competing set and $b^{m}_{q(s,t)}(O_{st})$ is the probability for observing the t^{th} test vector of s^{th} sub-band in the $q(s,t)$ state of the m^{th} competing speaker models. In order to obtain the required state sequences, the test utterance has to be time-aligned with the sub-band models of each competing speaker using the Viterbi algorithm and then the backtrack procedure has to be applied.

It should be noted that in order for the above weighting scheme to be meaningful, the corresponding states in the sub-band models of the target speaker and each of the competing speakers have to represent equivalent acoustic events. This equivalency can be encouraged during the training procedure by using the speaker independent sub-band models to initialise or *seed* the training of all required sets of the sub-band models.

The main attraction of the adopted approach for choosing the competing speaker models is its superior ability in reducing the false acceptance error [2]. This is because when the test utterance is produced by an impostor, the competing speaker models will be close to the test contour and not necessarily to the target model. As a result, $b_{sj}(O_{st})$ and $[w_t(s)]^{-1}$ both will become small and thereby the probability of false acceptance will be reduced significantly

4. SPEECH DATA AND FEATURES

The speech data used for this study was a subset of the BT Millar speech database [3]. 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, which was recorded in a quiet environment, 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 (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 [6]. 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.

5. EXPERIMENTAL INVESTIGATION

For the purpose of experiments an adverse effect was simulated by contaminating 1/3 of the test utterances with a narrow band noise (0-600 Hz). The HMM configuration used was a four state left-to-right structure without the "skip" transition and two Gaussian mixture per state. The first set of experiments were conducted using the SB-MFBO features. The results of this study are presented as a function of SNR in Figure 1. In order to perform a meaningful comparison, the figure also includes the results obtained for three other techniques in a similar experimental condition. These methods are the conventional full-band HMM (FB-HMM), FB-HMM with unconstrained cohort normalisation (FB-HMM+UCN) [2] and sub-band HMMs with SNR based recombination weights (SNR-RW).

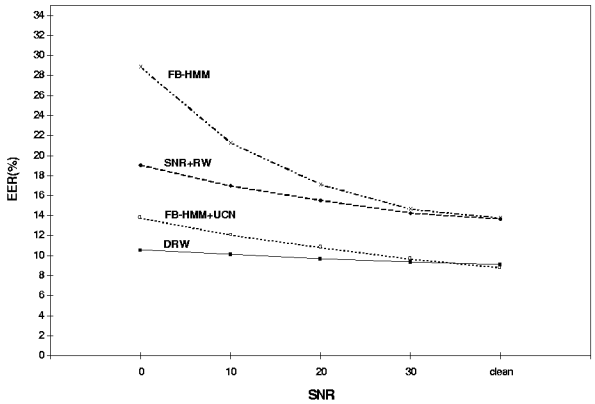

Figure 1 : Relative performance of the considered full- and sub-band approaches as a function of SNR for SB-MFBO features.

The robustness of DRW is clearly evident from these results. The figure also shows the benefit of using the score normalisation in the full-band approach. Moreover, SNR-RW appears to work reasonably well in compensating the effect of interference. This may be expected because the contamination here is due to additive noise. However, it should be emphasised again that this method, unlike DRW or UCN, cannot be used in tackling the effects of various other forms of interference.

The above experiments were repeated using SB-MFCCs. The results of this investigation are presented in Figure 2. As before, the DRW method exhibits a relatively flat response across the considered SNR range. However, the overall performance of FB-HMM+UCN is noticeably better than that of the DRW approach. The reason for this must be 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 relatively uncorrelated features for individual sub-bands. However, this can also result in a more detailed representation of the overall spectral envelope variations [10]. For example, in the case of four sub-bands, the detail of the overall spectral variations measured by the 1^{st} and 2^{nd} DCT basis functions of the different sub-bands are rather similar to that of 4^{th} and 8^{th} DCT basis functions of the full-band respectively (Figure 3). This

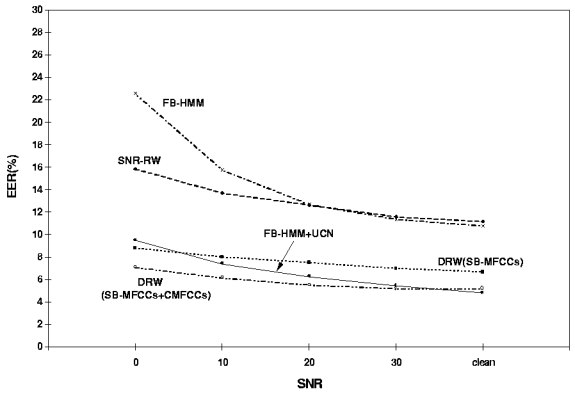

Figure 2 : EER for different approaches as a function of SNR using SB-MFCCs / SB-MFCCs and CMFCCs.

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.

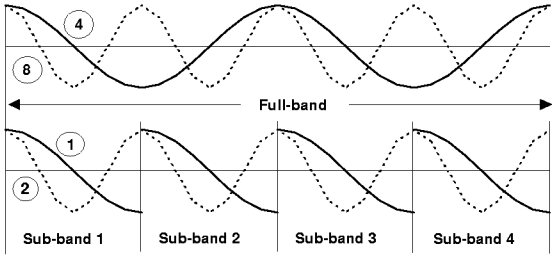

Figure 3 : Comparison between the full and sub-band DCT basis functions.

A method to tackle this problem is to generate an additional model for the target speaker and also for each of the competing speakers using MFCCs that are not represented by SB-MFCCs. Since these features are complementary to SB-MFCCs they are referred to as CMFCCs. In this case, the SBV algorithm has to be modified by replacing the term

$$S^{-1}\sum_{s=1}^{S}\log\left(w_t(s)b_{st}(O_{st})\right)$$

in equations (1) and (2) with

$$\alpha S^{-1}\left(\sum_{s=1}^{S}\log\left(w_t(s)b_{st}(O_{st})\right)\right)+(\alpha-1)\log\left(w_t'(s)b_t'(O_t')\right)$$

where, α is a combination factor between 0 and 1, O_t' is the t^{th} CMFCC vector of the test utterance, $b_t'(O_t')$ is the probability for observing O_t' in the j^{th} state of the CMFCC based model of the target speaker, and $w_t'(s)$ is a weighting factor which is computed using the CMFCC based competing speaker models. The use of these weights provides the possibility of correcting each frame level score 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 α = 0.5) presented in Figure 2 imply that the gains which can be achieved through the use of these full-band features in the above manner are more significant.

6. CONCLUSIONS

A new method for determining the recombination weights in sub-band based speaker verification has been presented. The approach is based on the use of background speaker models and aims to reduces the effect of the mismatch between the band-limited segments of test utterance and the corresponding sections in the target speaker model. The effectiveness of this 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 certain 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 mismatch which may be due to the time- and/or frequency-localised anomalies.

7. REFERENCES

[1] Allen J.B., "How do humans process and recognize speech ?," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 567-577, 1994.

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

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

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

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

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

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

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

[9] Lockwood P. and Boudy J. "Experiments with a non-linear spectral subtractor (NNS), hidden Morkov models and the projections, for robust speech recognition in car," *Speech Com.,* vol. 11, pp. 215-228, 1992.

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