THE USE OF SUB-BAND CEPSTRUM IN SPEAKER VERIFICATION

P. Sivakumaran and A. M. Ariyaeeinia

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

ABSTRACT

This paper focuses on the spectral representation of the sub-band cepstrum in relation to that of the full-band cepstrum. Through theoretical analysis it is shown that the net spectral information content of the cepstral coefficients with the same index in different sub-bands is only comparable to that of a full-band cepstral parameter whose quefrency is given by the product of that specific index with the number of sub-bands. A new method is proposed to tackle this deficiency of the sub-band cepstrum when it is used in the context of text-dependent speaker verification. The experimental investigations have clearly demonstrated the effectiveness of this method in speaker verification.

1. INTRODUCTION

In the conventional speech feature extraction process, each feature vector is generated by utilising the entire frequency spectrum of a given speech frame. Therefore, when the speech signal is partially degraded by an anomaly which is localised in time and frequency, the feature vectors that are generated within the time-span of that anomaly are completely contaminated. In such cases, however, it is likely that the unaffected parts of the spectral regions contain useful information for speaker discrimination. A logical way to tackle this problem is to split the entire frequency domain into a number of sub-regions and to use the spectral information contained in each of these regions to generate independent feature vectors. This technique is commonly known as the sub-band analysis and has been studied in the context of both speech and speaker recognition [3][4][7][10].

The use of sub-band analysis is also motivated by the fact that it closely resembles the front-end processes involved in human perception [1]. Moreover, the technique provides a way to emphasis the sub-bands that are more specific to the speaker and gives the possibility of relaxing the conventional time-synchrony assumption between the sub-bands [4].

Cepstrum has been the predominant speech feature type in both speech and speaker recognition. It therefore seems natural that the sub-band analysis is incorporated in the methods for extracting this type of feature [3][4][7][10]. This paper takes a closer view of the sub-band cepstrum in the context of text-dependent speaker verification.

The paper is organised in the following manner. The next section provides a review of the sub-band based speaker verification systems operated in the text-dependent mode and discusses various related issues. Section 3 focuses on the sub-band cepstral parameters. Section 4 gives a description of the utilised speech database. The experimental work and results are detailed in Section 5, and the overall conclusions are presented in Section 6.

2. SUB-BAND BASED TEXT-DEPENDENT SPEAKER VERIFICATION

In a typical sub-band based speaker verification method, each registered speaker is represented using a set of reference models in which each model is formed using the feature vectors of a particular sub-band. A simple strategy for text-dependent verification is to independently time-align the given sub-band vector sequences to the corresponding reference models, and to use the resulting scores to make the final decision. However, the time-warping paths obtained in this manner are less reliable because a sub-band vector consists of less spectral information than that of the full-band.

A possible method to tackle this problem is to recombine the intermediate outcomes of the separate time-alignment processes at certain pre-defined stages. In theory, it would be ideal to set each of these recombination stages to correspond to a *certain time segment*, such as a phoneme, syllable or word [4]. This will ensure the time-resynchrony of the speech events in different sub-bands at recombination and thereby will prevent the need to reintroduce the time-synchrony assumption between sub-bands. In practice, however, it has been found that the recombination stages set on this criterion has performed poorly compared to the simple frame-level recombination [10]. This may be due to the fact that many certain time segments are relatively too long in duration and thus incapable of limiting the extensive use of partial information. Another problem is in reliably defining the boundaries of the certain time segments. In this study it was decided to choose the simple frame-level recombination.

Another critical issue in the sub-band based speaker verification is the recombination process itself. Ideally, for this purpose, the scores of different sub-bands are to be fused in a constructive way so that sub-bands that are specific to the target speaker are emphasised while the contaminated ones are de-emphasised or removed. The main step in accomplishing this is to determine a weight for each sub-band score. In the literature, there have been a number of proposals for determining these weights [4][10]. A brief description of each of these techniques and an analysis of their strengths and weaknesses are given below.

One method to compute the required weights is to use the a priori knowledge of the sub-bands effectiveness in speaker discrimination. This knowledge may be gained simply through a series of experiments using a given set of speech data [10]. However, a more formal method is through discriminative training [4]. In general, this technique is 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

0-7803-6293-4/00/$10.00 ©2000 IEEE.

(produced by the true speaker) is contaminated in the regions where the weights are relatively high, then the approach can lead to an increase in the false rejection error. An obvious way to tackle this problem is to incorporate an estimated level of contamination of the test utterance in the process of generating the weights.

If the contamination is due to additive band-limited noise, then the recombination weights may be computed as SNR dependent [4]. 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 this technique, it has to be assumed that the interfering noise remains stationary during speech activities. Obviously, this cannot be the case in many practical applications. To tackle this problem a technique has been introduced [8] which 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 of this technique is that, for accurate estimation of the noise level, a relatively large speech segment (typically in the range of 1-2 s) is required.

In practice, additive band-limited noise is one of several types of anomalies that cause the sub-band contamination. Other such anomalies include those resulting from speaker generated variations and changes in the environmental and transmission channel conditions. An effective method to tackle the problems caused by time and frequency localised anomalies in the sub-band technique has been proposed by the authors in an earlier study [10]. This technique is referred to as dynamic recombination weight (DRW) and is based on the use of a set of background speaker models capable of competing with the sub-band model set of the target speaker. The competing speaker model set can be selected based on its closeness to either the target model set or the test utterance. For the purpose of this study, the second approach was chosen because of its superior ability in reducing the false acceptance error [2].

As reported in [10], this technique can be incorporate into the HMM framework by modifying the Viterbi algorithm as follows:

Step 1 : Initialisation :

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

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

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

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

Step 3 : Termination : final score

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

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, J 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 of the form

$$\log w_t(s) = -\frac{1}{L}\sum_{l=1}^{L}\log b^{l}_{q(s,t)}(O_{st}) \tag{5}$$

where L is the number of speakers in the selected competing set and $b^{l}_{q(s,t)}(O_{st})$ is the probability for observing the t^{th} test vector of s^{th} sub-band in the $q(s,t)$ state of the l^{th} competing speaker models.

3. SUB-BAND CEPSTRUM

As noted in the introduction, the cepstrum is the most commonly used type of speech feature in sub-band analysis. This section focuses on the generation of various types of sub-band cepstral features in order to determine their spectral representations in relation to that of the corresponding type of full-band cepstral parameters.

The first step in generating any type of sub-band cepstral features is the computation of the magnitude spectrum. In order to accomplish this, the utterances are usually pre-emphasised using a first-order digital filter. Each utterance is then segmented into fixed size frames at predetermined intervals using a Hamming window, and subjected to a fast Fourier transform (FFT).

One method of determining the sub-band cepstral coefficients is first to compute the logarithm of the magnitude spectrum and then group them according to the preset frequency divisions. The required cepstral parameters are obtained by independently applying the inverse FFT (IFFT) in each of these groups. Since the logarithm of the magnitude spectrum is real and even, the IFFT formula to compute the sub-band cepstral coefficients can be expressed in the following form:

$$c^{cb}(s,n) = \frac{1}{(N/S)}\sum_{k=(s-1)N/S}^{(sN/S)-1} Y(k)\cos\left(\frac{2\pi nk}{N/S}\right) \qquad s = 1,2....S \tag{6}$$

where $c^{cb}(s,n)$ is the n^{th} cepstral coefficient of the s^{th} sub-band, N is the number of log spectral magnitudes in the full-band and $Y(k)$ is the k^{th} log spectral magnitude. In this formulation it is assumed that the frequency range is divided into S, non-overlapping, equally spaced sub-bands.

From the log spectral magnitudes the full-band cepstral coefficients can be computed using the following expression:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} Y(k)\cos\left(\frac{2\pi nk}{N}\right) \tag{7}$$

By comparing equation (6) and (7), it can be seen that

$$c(Sn) = \frac{1}{S}\sum_{s=1}^{S} c^{cb}(s,n) \tag{8}$$

which indicates that the mean value of the n^{th} cepstral coefficient of a S sub-band system is equal to the $(Sn)^{th}$ cepstral coefficient of the full-band. For example, in a 4 sub-band system, the mean values of the first three cepstral coefficients are equal to that of the 4^{th}, 8^{th} and 12^{th} full-band cepstral coefficients respectively.

The cepstral parameters obtained in the above manner are referred to as the *real cepstrum* [6]. This is because the method uses only the magnitude information of the spectrum and ignores the phase information.

Another way of generating sub-band cepstral coefficients is first to analyse the magnitude spectrum using a mel-scale filterbank [5]. The log-energy outputs of the filterbank $\{Y'(k), k = 0,1,..,N'-1\}$ are then grouped according to the preset frequency divisions. In this case, it is common to use the discrete cosine transform (DCT) instead of IFFT to determine the cepstral coefficients [5]. The DCT based formulae to compute the sub-band and full-band cepstral coefficients are

$$c^{cb}(s,n) = \frac{\cos(\pi n(s-1))}{(N'/S)} \sum_{k=(s-1)N'/S}^{(sN'/S)-1} Y'(k) \cos\left(\frac{\pi n(k+0.5)}{N'/S}\right), \text{ and} \quad (9)$$

$$c(n) = \frac{1}{N'} \sum_{k=0}^{N'-1} Y'(k) \cos\left(\frac{\pi n(k+0.5)}{N'}\right) \quad (10)$$

respectively. In this case, only the mean value of each even numbered sub-band cepstral coefficient is equal to that of a full-band parameter, i.e.

$$c(2Sn) = \frac{1}{S} \sum_{s=1}^{S} c^{cb}(s,2n) \quad (11)$$

In the case of odd numbered cepstral coefficients, the DCT basis function that operates in every even numbered sub-band will be negated in relation to that which operates in the corresponding portion of the full-band (Figure 1). Thus the relationship in equation (11) does not apply here. However, by considering the spectral variation measured by the DCT basis function, it can be argued that the spectral information represented by $c(S(2n-1))$ is the same as that of $(1/S)\sum_{s=1}^{S} c^{cb}(s,(2n-1))$.

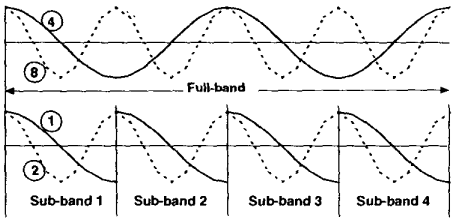

Figure 1: Examples of full- and sub-band DCT basis functions.

The cepstral parameters computed in the above manner are commonly referred to as mel-frequency cepstral coefficients (MFCCs). In order to maintain consistency with an earlier investigation conducted by the authors [10], it was decided to use this type of feature in the experimental work. In generating these features, each utterance was segmented into 32 ms frames at intervals of 16 ms using a Hamming window, and subjected to an 8th order FFT.

Another type of cepstral parameter commonly used in sub-band analysis is the LPC-derived cepstrum. In the computation of these parameters, the spectrum in each sub-band is modelled independently using an all-pole filter. The technique to accomplish this is known as *selective linear prediction* [9]. There exists an efficient recursive method to compute the required cepstral coefficients from the parameters of the modelling filter [6]. Although this method inherently uses both the magnitude and the phase information of the spectrum, the result is effectively a scaled version of the *real cepstrum* [6]. This is due to the fact the all-pole filters used in this case are of minimum phase. Although this may imply that equation (8) can be applied here, this is not necessarily true. The reason for this is that the all-pole spectral fit for each sub-band is performed independently and thus it is difficult to tune them to match that of an all-pole fit of the full-band spectrum exactly.

From the above discussions it is evident that in a given sub-band system, the local cepstral features cover only a portion of the spectral information represented by the full-band cepstral parameters. More specifically, the net effect of cepstral coefficients with identical indices in different sub-bands is only equal to that of a full-band cepstral parameter whose quefrency is given by the product of that specific index with the number of sub-bands.

One method for tackling this problem is to supplement the sub-band cepstral coefficients with the full-band parameters that are not covered by them. For example, in the case of a 4 sub-band system, the full-band features $c(n)$, n = 1-3,5-7,9-11, may be supplemented to the sub-band features. Of course, this prevents the complete realisation of the benefits of the sub-band analysis. It has, however, been shown that using sub-band and full-band features in this manner can lead to a better speaker verification accuracy than that obtainable using any of these individually [10].

In order to tackle this problem more effectively a new method is proposed here which involves the use of the cepstral parameters generated from a set of different sub-band systems. For example, it is possible to cover a large part of the spectral information represented by the full-band cepstral parameters 1-12, if the cepstral coefficients from sub-band systems 2-4 are utilised. In this case, the Viterbi algorithm given by equations (1)-(4) could be modified by replacing the term

$S^{-1}\sum_{s=1}^{S} \log(w_t(s)b_{st}(O_{st}))$ in equations (1) and (3) with

$M^{-1}\sum_{m=1}^{M}\left[\alpha_m S_m^{-1}\sum_{s=1}^{S_m} \log(w_{mt}(s)b_{mst}(O_{mst}))\right]$, where $\sum_{m=1}^{M}\alpha_m = 1$ {in this study α_m is simply set to $1/M$}, M is the utilised number of sub-band systems and the subscript m indicates the association of the m^{th} sub-band system. This method certainly provides more flexibility in dealing with time and frequency localised anomalies. Its main drawback, however, is the increase in computational complexity. For the purpose of this study the above two methods are simply referred to as modified sub-band analysis 1 & 2 (MSBA1 and MSBA2) respectively.

4. SPEECH DATA

The speech data used for this study was a subset of the BT Millar speech database [3][10]. The subset consisted of 25 repetitions of digit utterances zero to nine spoken by 20 male native English 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.

5. EXPERIMENTAL INVESTIGATION

All the experiments were conducted within the HMM framework. The HMM topology used was a four state left-to-right structure without the "skip" transition and with two Gaussian mixtures per state. The first set of experiments was carried out to determine

the relationship between the performance of SB-MFCCs (here SB stands for sub-band) and the number of sub-bands. As an exception, in this part of the investigation, the recombination process is performed at the word level with unity weights. Figure 2 presents the results of this study. For reference purpose, this figure also includes the results obtained for SB-MFBOs (MFBO stands for mel-scale filterbank outputs) under the same experimental conditions.

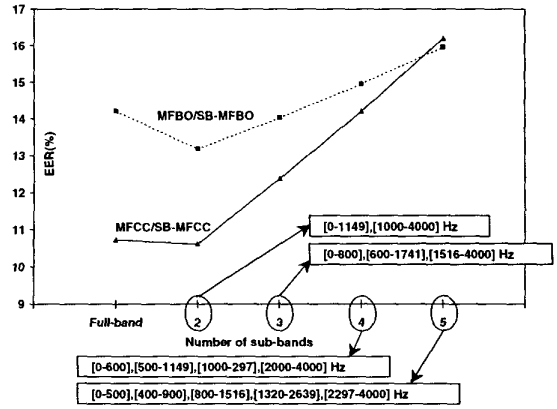

Figure 2: Equal error rates (EERs) as a function of the number of sub-bands for two types of speech features (for each sub-band group, the frequency range of each individual sub-band is also shown).

These results indicate that the two sub-band system is capable of achieving better performance than that of the conventional full-band system. A consistent increase in the verification error rate is observed in both cases when the number of sub-bands is increased from 3 to 5. This increase is relatively larger in the case of SB-MFCCs. In fact, it is seen that in the case of the 5 sub-band system, SB-MFBOs outperform SB-MFCCs. This implies that the reduction in the spectral information due to narrowing frequency bands is higher in SB-MFCC.

For the next part of the experimental investigation an adverse effect was simulated by contaminating 1/3 of the test utterances with a narrow band noise (0-600 Hz). The aim of this part of the investigation was to evaluate the relative performance of the conventional sub-band cepstrum based analysis (SBCA) and the modified versions MSBA1 & MSBA2. In the case of MSBA1, a 4 sub-band system was chosen whereas in the case of MSBA2, sub-band systems 2-4 were used. In all three cases DRW was applied. The results of this study are presented as a function of SNR in Figure 3. In order to perform a meaningful comparison, the figure also includes the results obtained for three other techniques used in similar experimental conditions. 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). These results are clearly in favour of MSBA2 with DRW.

6. CONCLUSIONS

It has been shown in this paper that the net spectral information of the cepstral coefficients with identical indices in different sub-bands is only comparable to that of a full-band cepstral parameter

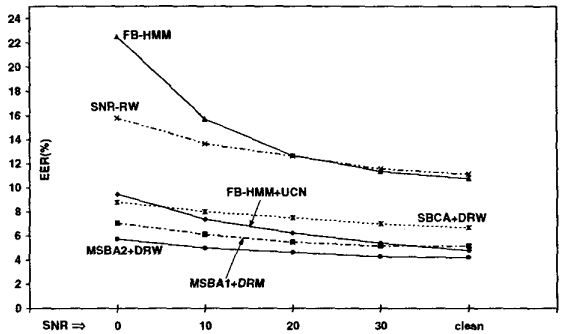

Figure 3: EER for different approaches as a function of SNR.

whose quefrency is given by the product of that specific index with the number of sub-bands. A new method is proposed to tackle this deficiency of sub-band cepstrum when it is used in the context of text-dependent speaker verification. The approach is based on the use of the cepstral parameters generated from a set of different sub-band systems. In the first part of the experimental investigation, the relationship between the size of the frequency bands and the spectral information content of sub-band cepstrum has been analysed. In the second part, the effectiveness of the proposed technique for speaker verification has been clearly demonstrated. The current work in this area includes further investigation into methods for tackling problems associated with the sub-band cepstrum.

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] 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.

[5] 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.

[6] Deller, J.R. *et al.*, *Discrete-Time Processing of Speech Signals*, Macmillan Inc., New York, 1993.

[7] Hermansky H., *et al.*, "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] Makhoul J., "Spectral linear prediction: Properties and application". *IEEE Trans. on ASSP*, vol. 23, pp. 283-296, June 1975.

[10] Sivakumaran P., Ariyaeeinia A.M. and Hewitt J.A. "Sub-band based speaker verification using dynamic recombination weights". *Proc. ICSLP'98*, vol. 3, pp. 551- 554.