MULTIPLE SUB-BAND SYSTEMS FOR SPEAKER VERIFICATION

P. Sivakumaran, A. M. Ariyaeinia and J. A. Hewitt
University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB, UK
{p.sivakumaran, a.m.ariyaeinia, j.a.hewitt}@herts.ac.uk

ABSTRACT

It is well known that the spectral information contained in the full-band cepstral parameters within a certain range is highly useful for speaker discrimination. However, this information cannot be fully represented by using the cepstral parameters generated in a single sub-band system. This paper focuses on methods to tackle this deficiency of the sub-band cepstrum in the context of text-dependent speaker verification. In particular, it focuses on a technique based on the collective use of the cepstral vectors generated from a set of different sub-band systems. The paper also addresses the procedure for an effective incorporation of this technique into the hidden Markov model (HMM) framework. Furthermore, the implementation issues are discussed and details of the experimental evaluation are presented.

1. INTRODUCTION

The sub-band analysis is the process of splitting the entire frequency domain into sub-regions and to use the spectral information contained in each of these regions to generate an independent set of feature parameters. In recent years, various research groups have shown interest in using this technique for speaker recognition [2,3,9,10]. The main motivation for this is the possibility provided by the sub-band analysis to emphasize the spectral regions that are specific to the target speaker and to deemphasize the ones that are contaminated.

One of the key issues to be addressed in designing any sub-band based speaker recognition system is the selection of the type of feature vectors. It may be thought that the cepstrum is the most obvious choice for this purpose. This is because the full-band cepstrum has been repeatedly shown to be the most effective feature type for speaker discrimination [7]. In the sub-band level, however, the cepstrum has an intrinsic deficiency in representing the required spectral information [10]. This paper attempts to tackle this problem in the context of text-dependent speaker verification. In particular, it focuses on an approach which is based on the collective use of the cepstral vectors generated from a set of different sub-band systems. In the remainder of this paper the above approach is referred to as multiple sub-band systems - MSBS.

The paper is organized in the following manner. The next section provides a review of the sub-band based text-dependent speaker verification systems. Section 3 discusses the fundamental problem in using the sub-band cepstral parameters. Section 4 focuses on the MSBS technique and the related issues. Section 5 gives a description of the utilised speech database, and the method used for the extraction of sub-band features. The experimental work and results are detailed in Section 6, and the overall conclusions are presented in Section 7.

2. SUB-BAND BASED TEXT-DEPENDENT SPEAKER VERIFICATION

In a typical sub-band based text-dependent 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 [9]. With such speaker modelling, a simple strategy for verification trial is first to time-align the feature vector set in each given sub-band to the corresponding reference model independently. The resulting scores associated with individual sub-bands can then be used to make the final decision. However, since the individual feature vector sets used in the process represent sub-spectral information, the time warping paths obtained in this manner may not be as reliable as that based on using full-band feature vectors.

A possible solution to this problem is to recombine the intermediate outcomes of the separate time-alignment processes at certain pre-defined stages [4][9]. In conventional systems, each of these recombination stages is set to correspond to the end of a certain time segment, such as a phoneme, a syllable or a word. This assures the time-synchrony of the speech events in different sub-bands. In practice, however, it has been found that the recombination stages set on this criterion performs poorly compared to the simple frame-level recombination [9]. 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. An additional problem is that of defining the boundaries of certain time segments reliably. In this study it was decided to choose the simple frame-level recombination.

3. SUB-BAND CEPSTRUM

If a $S$ sub-band system is formed by equally distributing the log spectral parameters $Y(k)$, $k = 0, 1, \ldots, K-1$ then the $p^{th}$ cepstral coefficient of the $x^{th}$ sub-band can be computed according to the following equation.

$$
c_x(s,p) = \frac{1}{K/S} \sum_{k=r+iK}^{(a_{n+1})} Y(k) \cos \left( \frac{a_{n} p (k + a_{s})}{K/S} \right) 
$$

where the values of $a_{1}$ and $a_{2}$ are either 2 and 0 or 1 and 0.5 respectively. The former set of values is commonly used in computing fast Fourier transform (FFT) (or linear predictive coding (LPC)) based cepstral parameters [5][6]. In this case, equation (1) implies that

$$
c(s) = \frac{1}{S} \sum_{x} c_x(s,p)
$$

where $c(s) \{ = c_{1}(s,i) \}$ is the $x^{th}$ full-band cepstral parameter. The latter set of values is commonly used in determining mel-frequency cepstral coefficients (MFCCs) [5]. Here, the relation-
ship between the full and sub-band cepstral parameters is of the following form

$$c(S_p) = \{1/S\} \sum_{r=0}^{N} (-1)^{p-r} c_r(s, p)$$

(3)

A common conclusion from equation 2 and 3 is

$$I(c(S_p)) = \sum I(c_r(s, p))$$

(4)

where I(x) is the spectral information contained in x. In other words, the net spectral information of the cepstral coefficients with identical indices in different sub-bands is only equal to that of a full-band cepstral parameter whose index is given by the product of that specific index with the number of sub-bands.

This implies that a sub-band system on its own cannot achieve the performance of the full-band system when the effects of frequency-localised spectral contaminations are negligible. This is because studies have already shown that, under such conditions, all the full-band cepstral parameters within a certain range are important for speaker verification [7].

One method to tackle this problem is to use the sub-band cepstral coefficients together with the full-band parameters that are not covered by them. For example, in the case of a 2 sub-band system, the full-band features c(p), p = 1, 2, 4, 5, 7, 8, 10, 11 may be supplemented with 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 [9]. In this paper, the above method is referred to as "X-sub-band system with full-band cepstral supplement - XSSFSCS". A more effective way to deal with the above problem is discussed in the next section.

4. MULTIPLE SUB-BAND SYSTEMS

The idea here is to use the cepstral vectors generated from a set of different sub-band systems. For example, the full-band cepstral parameter in the range p to p + q, i.e.,

$$\{c(p), c(p + 1), \ldots, c(p + q)\}$$

can be spectrally represented by the 1st cepstral coefficients of $$S(0, 1)\ldots S(q-1)$$ and $$S(q)$$, where $$S(r)$$ is the sub-band system which consist of $$p + i$$ frequency bins (or sub-bands), i.e.,

$$\{c_p(1,1), c_p(2,1), \ldots, c_p(p,1), c_{p+1}(1,1), c_{p+1}(2,1), \ldots, c_{p+1}(p+1,1), \ldots, c_{p+q}(1,1), c_{p+q}(2,1), \ldots, c_{p+q}(p+q,1)\}$$

Such a representation may, however, need to include systems of unusually large number of sub-bands. This in turn may detract from the reliability of the speaker verification process. An alternative representation would be based on using R ≪ q sub-band systems in which certain sub-band systems contribute more than one set of cepstral coefficients. In this approach, it is possible to avoid the systems that have unusually large number of sub-bands. However, the difficulty is in determining the sub-band representation for c(x) (p ≪ q) when x is not divisible by any of S, S, \ldots, S where S is the number of frequency bins in the pth sub-band system. It should be noted that such cases are usually small in number and therefore supplementing the full-band parameters to that of the sub-band does not seriously obstruct the realisation of the benefits of the sub-band analysis.

In this study, in order to incorporate the multiple sub-band systems approach into the hidden Markov model (HMM) framework, each registered speaker is represented using $$\lambda_o$$, where $$\lambda_o$$ is the N-state, M-mixture, left-to-right HMM associated with the $$i^{th}$$ frequency bin of the $$j^{th}$$ sub-band system.

In this framework, the key task is the estimation of the model parameters using a set of L training utterances, $$O'_{i=1} \ldots L$$, where

$$O' = \{O_{i} = \{o_{ij} \}_{i=1 \ldots l, j=1 \ldots q} \}_{i=1 \ldots L, j=1 \ldots q}$$

and $$o_{ij}$$ is the set of cepstral coefficients chosen from the $$i^{th}$$ frequency bin in the $$j^{th}$$ sub-band system of the $$l^{th}$$ vector sequence. In order to accomplish this, a modified version of the Baum-Welch re-estimation procedure is used. In this approach, for a given training utterance $$O'$$, the probabilities $$\gamma'_{i}(i, j)$$ and $$\gamma'_{j}(j, m)$$ are assumed to be equal for all the HMMs used to represent the target speaker and are computed collectively. $$\xi_{i}(i, j)$$ is the probability of being in state i at time t, and state j at time t+1, and $$\gamma'_{j}(j, m)$$ is the probability of being in state j using $$m^{th}$$ mixture component. The reasons for this are to improve the reliability of the model parameters and to force a frame-level sub-band recombination.

Suppose the parameters of the HMM in the $$i^{th}$$ frequency bin of the $$j^{th}$$ sub-band system are denoted as follows:

$$A = \{a_{ik}^{o} \}_{i=1 \ldots N, k=1 \ldots M}$$

$$C = \{c_{i}^{o} \}_{i=1 \ldots N}$$

$$\mu^{o} = \{\mu_{i}^{o} \}_{i=1 \ldots M}$$

$$U^{o} = \{U_{i}^{o} \}_{i=1 \ldots M}$$

where $$a_{ik}^{o}$$ is probability of transition from state i to j, and the parameters $$c_{i}^{o}$$, $$\mu^{o}$$ and $$U^{o}$$, which are associated with the $$m^{th}$$ mixture in state j, are the weight, p-dimensional mean vector and pvp covariance matrix respectively. The re-estimation formulas used in the model training can be expressed as follows:

$$a_{ik}^{o} = \sum_{t=1}^{T} \sum_{j=1}^{N} \pi_{i}^{t+1}a_{ik}^{t}$$

$$c_{i}^{o} = \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_{i}(i, j)\pi_{i}^{t}$$

$$\mu^{o} = \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_{i}(i, j)\pi_{i}^{t}\mu_{i}$$

$$U^{o} = \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_{i}(i, j)\pi_{i}^{t}(\mu_{i}^{o} - \mu_{i}^{o})(\mu_{i}^{o} - \mu_{i}^{o})$$

where

$$\gamma'_{i}(i, j) = a_{i}(i)a_{j}(j)/(\sum_{k=1}^{M} a_{i}(i)a_{k}(j))$$

and ' denotes the transpose operation.

In equation (9), the values $$a_{i}(i)$$ and $$\beta_{j}(j)$$ are the forward and backward probabilities associated with the $$l^{th}$$ vector sequence respectively. For the purpose of text-dependent speaker verification, the forward and backward probabilities are commonly computed using the following induction formulas.

$$a_{i}(i) = \sum_{t=1}^{T} \alpha_{t}(t)$$

(11)

with the initial conditions

$$\alpha_{t}(t) = \begin{cases} \alpha_{i}(i), & t = 1, 2, \ldots, T - 1 \\ \beta_{t}(t), & t = 2, 3, \ldots, N \end{cases}$$

where

$$\alpha_{i}(i) = \begin{cases} \alpha_{i}(i), & t = 1, 2, \ldots, T - 1 \\ \beta_{t}(t), & t = 2, 3, \ldots, N \end{cases}$$

and

$$\gamma'_{j}(j, m) = \begin{cases} \gamma'_{j}(j, m), & t = 1, 2, \ldots, T - 1 \\ \beta_{t}(t), & t = 2, 3, \ldots, N \end{cases}$$

with the initial conditions
\[ a'_i(t) = b_i(O'_i) \] \[ a''_i(i) = 0, \ i = 2, 3, ..., N. \] (12)

\[ b'_i(i) = \sum_{n=1}^{N} a_n b_n(O'_i) \beta'_n(j), \quad i = T-1, T-2, ..., 1 \]
\[ \beta'_n(i) = 1, \ i = 1, 2, ..., N. \] (13)

The computation of both forward and backward probabilities involves a large number of multiplications of numbers less than unity which can cause arithmetic underflow conditions in practice. In order to prevent this, the scaling procedure described in [8] can be adopted. In this case, the probability of generating \( O'_i \) by the model set \{ \[ \theta_n \] \} \( i = 1, 2, ..., N \) and \( P_n \) (which is used in equations 9 & 10) can be determined using the associated scaling factors [8].

In the above equations, \( b_i(O'_i) \) represents \( \sum_{n=1}^{M} c_{n} b_n(O'_i) \) and the probability \( \beta_n(O'_i) \) is estimated in the following manner:

\[ \log \beta_n(O'_i) = \frac{1}{R} \sum_{r=1}^{R} \sum_{s=1}^{S} \log h'_n \mathcal{Q}_n (\alpha'_n; \mathbf{p}^n, U^r_n) \] (15)

where \( \mathcal{Q}_n \) is a \( p \)-dimensional Gaussian density function with the mean vector \( \mathbf{p}^n \) and the covariance matrix \( U^r_n \). \( b'_n \) and \( h'_n \) are the weights that control the contribution of different sub-band systems and their frequency bins respectively. In this study, for the purpose of training, \( g'_r \) is set to 1 and \( h'_n \) is chosen based on a priori knowledge of the sub-band performance [9].

Given the test utterance \( \{ O = (\alpha^r) \}_{r=1, 2, ..., R} \) and the target speaker models \( \{ \hat{b}_n \}_{n=1, 2, ..., N} \), a verification score can be obtained using the Viterbi algorithm in the following manner.

**Step 1: Initialisation**

\[ \delta_1(i) = \frac{1}{R} \sum_{r=1}^{R} \hat{b}_i \sum_{s=1}^{S} \log h'_n \mathcal{Q}_n (\alpha'_n; \mathbf{p}^n, U^r_n) \] (16)

for \( j = 2 \) to \( N \), \( \delta(j) = -\infty \) (17)

**Step 2: Main Recursion**

for \( t = 2 \) to \( T \) and \( j = 1 \) to \( N \)

\[ \delta(t) = \max_{l \in \text{h}} \left[ \delta_{l}^{l-1}(i) + \log a_g + \frac{1}{R} \sum_{r=1}^{R} \hat{b}_i \sum_{s=1}^{S} \log h'_n \mathcal{Q}_n (\alpha'_n; \mathbf{p}^n, U^r_n) \right] \] (18)

**Step 3: Termination**

final score

\[ l = \max_{l \in \text{h}} \delta_{l}(i) \] (19)

where \( b'_n(\alpha'_n) = \sum_{n=1}^{M} c_{n} \mathcal{Q}_n (\alpha'_n; \mathbf{p}^n, U^r_n) \). (20)

In the verification phase, determining \( h'_n \) solely based on a priori knowledge of the sub-band performance is not appropriate. This is because the weights can lead to an increase in the false rejection error if a test utterance (produced by the true speaker) is contaminated in the regions where emphases are relatively high. 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 these weights may be computed as signal to noise ratio (SNR) dependent [9]. In practice, however, 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 [9]. This technique is based on the use of a set of background speaker models capable of competing with the model set of the target speaker [1]. Based on this approach, the required weights can be obtained as follows.

\[ \log h'_n = \left( \frac{1}{I} \right) \sum_{l=1}^{I} \log h'_n^{l}(\alpha'^r) \] (21)

where \( w^l \) is a weight which is set based on a priori knowledge of the sub-band performance, the superscript \( i \) indicates the association of the \( i \)th competing speaker model and \( I \) is the number of speakers in the selected competing set (which was set to 2 in the experiments). In order to obtain the required state sequences, \( q(t) = 1, 2, ..., T \), the test utterance has to be time-aligned with the model set of each competing speaker using the Viterbi algorithm and then the backtracking procedure has to be applied.

### 5. SPEECH DATA AND FEATURES

The speech data used for the evaluation experiments was a subset of the BT Milluar speech database [11][2]. The subset consisted of 25 repetitions of digit utterances one to nine and zero spoken by 63 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.

For the purpose of this study, MFCCs were chosen. 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 8th order FFT. The resulting energy spectrum for each frame was analysed appropriately using a mel-scale filterbank [5]. The frequency range was divided into bins according to the selected sub-band systems. The log-energy outputs of the filterbank were then grouped in relation to these frequency bins. The discrete cosine transform (DCT), which is implied by equation (1) when \( a_1 = 1 \) and \( a_2 = 0.5 \), was applied to each of these groups to generate the required sub-band MFCCs. The full-band MFCCs were generated by applying the DCT to the entire set of the log-energy outputs of the filterbank.

### 6. EXPERIMENTAL INVESTIGATION

In order to develop an efficient MSBS method, it is necessary to know the relative speaker discrimination ability of each individual full-band cepstral coefficients. Since the cepstral parameters are fairly uncorrelated, their relative speaker discrimination power can be given in terms of the ratio of inter- to intra-speaker variability [7]. Figure 1 shows this ratio, estimated using a telephone quality database (provided by BT Labs), as a function of the cepstral index. It can be seen that, as expected, the lower and higher order coefficients are less effective than the middle order ones. In particular, the cepstrum region 3-12 appears to have the highest speaker discrimination abilities and therefore chosen for the purpose of this study. All the parameters in the selected region can be spectrally represented by using appropriate cepstral coefficients of the sub-band systems 3-5 (Figure 2), except \( c(7) \) and \( c(11) \), which require the use of the sub-band systems 7 and 11 respectively. Since the preliminary experimental study have shown that the use of more than 6 frequency bins
in a sub-band systems can compromise the reliability of the speaker verification process, it was decided to supplement $c(7)$ and $c(11)$ to the chosen sub-band parameters.

![Graph](image1.png)

**Figure 1:** Estimated speaker discrimination ability of each individual cepstral coefficient.

![Graph](image2.png)

**Figure 2:** The parameter sets chosen from the utilised sub-band systems to spectrally represent the full-band cepstral coefficients that are found to be the most useful for speaker verification.

In order to experimentally investigate the effectiveness of the MSBS technique for speaker verification, an adverse effect was simulated by contaminating 1/3 of the test utterances with a narrow band noise (0-600 Hz). The HMM topology used throughout the experimental work was a four state left-to-right structure without the “skip” transition and two Gaussian mixture per state. The results of this study are presented in terms of the equal error rate as a function of the SNR in Figure 3. In order to have a meaningful basis for evaluating the performance of the method, the results obtained for five other techniques under identical experimental conditions are also presented in this figure. These techniques are the conventional full-band HMM (FB-HMM), FB-HMM with unconstrained cohort normalisation (FB-HMM+UCN) [1], 4-sub-band system with dynamic recombination weights (4SB-DRW) [9], 455FCS (Section 3), and 4-sub-band system with SNR based recombination weights (SNR-RW) [9]. These results clearly indicate the superior performance of MSBS under various noise level conditions.

**7. CONCLUSIONS**

The problem of not being able to represent the speaker specific information contained in the full-band cepstral coefficients by using

![Graph](image3.png)

**Figure 3:** EER for different methods as a function of the SNR.

the cepstral vectors generated from a single sub-band has been described. As an effective solution to this problem, the use of the cepstral vectors computed from a set of different sub-band systems has been presented and discussed. In order to incorporate this method into the HMM framework effectively, appropriately modified versions of the Baum-Welch and the Viterbi algorithms have been introduced. Based on the empirical knowledge of the relative speaker discrimination abilities of the full-band cepstral coefficients, an efficient implementation of the considered method has been presented. Finally, the effectiveness of this method has been demonstrated through a set of comparative experimental investigations.

**8. REFERENCES**