SUB-BAND BASED TEXT-DEPENDENT SPEAKER VERIFICATION

P. Sivakumaran, A. M. Ariyaeeinia** and M. J. Loomes***

*20/20 Speech Ltd., Malvern, Worcestershire, WR14 3SZ, UK

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

** p.sivakumaran@2020speech.com, **{a.m.ariyaeeinia, m.j.lommes}@herts.ac.uk*

** Sivakumaran was with the University of Hertfordshire during the course of this work*

Contact:

Aladdin M. Ariyaeeinia

University of Hertfordshire
Collage Lane
Hatfield
Hertfordshire
AL10 9AB
UK

Tel: +44 (0)1707 284348
Fax: +44 (0)1707 284899
E-mail: a.m.ariyaeeinia@herts.ac.uk

List of Unusual Symbols and Abbreviations Used

$c_s(s,p)$	p^{th} cepstral coefficient of the s^{th} sub-bands { $c_1(1,p) = c(p)$ is the p^{th} full-band cepstral parameter}
S	number of sub-bands
$Y(k)$	k^{th} log spectral magnitude
K	number of log spectral magnitudes
$Y''(k)$	k^{th} log-energy outputs of the mel-scale filterbank
K''	number of log-energy outputs of the mel-scale filterbank
h_t	weight associated with the t^{th} segment
U	number of competing speakers
$(1-\eta_s)$	average speaker verification error rate in the s^{th} sub-band
R, q	number of sub-band-systems
Superscript s	implies the association of the s^{th} sub-band
Superscript u	implies the association of the u^{th} competing speaker
Superscript r	implies the association of the r^{th} sub-band-system
$'$	implies the association of the complementary full-band cepstral parameters
SB-	Sub-band based
DBSW	dynamic band-limited segmental weights
MFBO	mel-scale filterbank output
SNRW	signal-to-noise ratio -based weighting factors
CFBCC	complementary full-band cepstral coefficient
SSFCS	sub-band system with full-band cepstral supplements
MSBSA	multiple sub-band-systems analysis
FWN	filtered white noise
RNT	real noise type

Number of pages: 50 (Text: 30 pages, Appendix: 3 pages, Figures and Tables: 17 pages)

Number of Figures: 15

Number of Tables: 1

Keywords: *Speaker verification*, *Sub-band analysis*, *Cepstrum*

Abstract

This paper addresses various issues involved in sub-band based text-dependent speaker verification. The first part of the discussions is concerned with the classification methods. An important issue addressed in this part is the determination of a set of weights which emphasises the sub-bands that are specific to the target speaker while de-emphasising or removing the contaminated ones. In particular, techniques for determining these weights dynamically according to the level of contamination in the sub-bands are described. Furthermore, the effectiveness of these methods is experimentally analysed through a set of comparative studies. The second part of the discussions focuses on the feature extraction process. Analytically, it is shown that for a sub-band system of S bands, the cepstral coefficients with the quefrency of p have a strong linear relationship to the $(S\times p)^{th}$ full-band cepstral parameter. With the aid of a set of experimental results, it is demonstrated that this means the conventional classification methods adapted to work with sub-band cepstral parameters may not be able to capture all the useful spectral information contained in the full-band cepstral parameters. In order to tackle this problem, two methods are described and their relative effectiveness is experimentally examined. The experimental investigations also include an examination of speaker discrimination abilities of different sub-bands and an analysis of different possible recombination levels.

1. Introduction

In the conventional 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][5][6][11][17-20]. The main attraction of this approach is that it provides the possibility of selectively de-emphasising the frequency spectral regions that are affected by anomalies. The other factors, which further motivate the use of the sub-band analysis in speaker verification, can be described as follows.

- The sub-band analysis closely resembles the front-end processes involved in human perception. The results of some psychoacoustics study have indicated that the human auditory mechanism decodes linguistic messages independently in different frequency sub-bands and the final decision is based on merging the information from these sub-bands [1].

- Different frequency spectral regions may not be equally effective for speaker verification. For example, it has been shown in [10] that, for some speakers, the upper part of the frequency spectrum is considerably more useful for the purpose of identification than the lower part.

- Transitions between more stationary segments of speech do not necessarily occur at the same time across different frequency bands [6]. This time-asynchrony may be due to a number of factors including the underlying asynchrony between different parts of the human speech production system and the asynchrony introduced by the communication channels [20]. Because of the use of complete synchronous feature vectors in conventional systems, the above effects are not accommodated. The sub-band based approach has the potential to relax this constraint and this should be investigated for the purpose of speaker recognition.

- Different verification strategies might ultimately be applied in different sub-bands. For example, feature vectors generated from different sub-bands can be selected to have different time/frequency resolutions [6].

This paper is concerned with the sub-band based text-dependent speaker verification systems. In such a system, 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 [17]. With this speaker modelling, a simple strategy for the 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 indi-

vidual 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 full-band feature vectors. A possible solution to this problem is to recombine the intermediate outcomes of separate time-alignment processes at certain pre-defined stages [6][11]. 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-resynchrony of the speech events in different sub-bands. Figure 1 shows the high-level operations involved in a typical sub-band based text-dependent speaker verification system.

In this study, Hidden Markov models (HMMs) are used for speaker representation. The reason for this choice is that HMMs provide a very effective and, in fact, the most commonly used framework for text-dependent speaker verification [16]. Furthermore, the investigations are based on the use of cepstral feature parameters. This is believed to be a natural choice as cepstrum is the predominant type of speech features in both speech and speaker recognition.

The paper is organised in the following manner. Section 2 describes the direct incorporation of the sub-band analysis into the HMM framework for text-dependent speaker verification. An important issue addressed in this section is the determination of a set of weights for emphasising the band-limited segments that are specific to the target speaker while de-emphasising or removing the contaminated ones. Three classes of methods for determining the required set of weights are proposed and their relative strengths and weaknesses are discussed. Section 3 is concerned with the sub-band cepstral parameters. It shows that the conventional verification methods adapted to work with sub-band cepstral parameters may not be able to capture all the useful spectral information contained in the full-band cepstral parameters. In order to tackle this problem, two new techniques are discussed in detail. Section 4 covers the experimental investigation carried out to evaluate the effectiveness of the proposed methods. It also includes two other related experimental studies: the usefulness of different sub-bands for speaker discrimination and the relative effectiveness of different possible recombination levels. The overall conclusions are presented in Section 5.

2. Sub-band Analysis in the HMM Framework

In order to directly incorporate the sub-band analysis into the HMM framework each registered speaker is represented by a set of HMMs in which individual models are formed in different sub-bands using an appropriate training algorithm [15]. In this case, the final verification score is obtained by modifying the Viterbi algorithm to that given below.

Step 1: Initialisation

From $s = 1$ to S

$$\delta_1^s(1) = \begin{cases} S^{-1}\sum_{s=1}^{S}\log\left(b_1^s(\mathbf{O}_1^s)h_1^s\right) & \text{If the initial state is a recombination level} \\ \log\left(b_1^s(\mathrm{O}_1^s)h_1^s\right) & \text{otherwise} \end{cases} \quad (1)$$

$$\text{From } j = 2 \text{ to } N \;\Rightarrow\; \delta_1^s(j) = -\infty \quad (2)$$

$$\psi_1^s(i) = 0 \quad (3)$$

Next s

Step 2: Main Recursion

From $t = 2$ to T,

From $j = 1$ to N,

From $s = 1$ to S

$$\delta_t^s(j) = \max_{1\le i\le N}\left[\delta_{t-1}^s(i) + \log\left(a_{ij}^s\right)\right] + \log\left(b_j^s(\mathbf{O}_t^s)h_t^s\right) \quad (4)$$

$$\psi_t^s(j) = \operatorname*{argmax}_{1\le i\le N}\left[\delta_{t-1}^s(i) + \log\left(a_{ij}^s\right)\right] \quad (5)$$

$$\text{If } t \text{ is the recombination level} \;\Rightarrow\; \delta_t^s(j) = S^{-1}\sum_{s=1}^{S}\delta_t^s(j) \quad (6)$$

Next s

Next j

Next t

Step 3: Termination

$$P = S^{-1}\sum_{s=1}^{S}\max_{1\le j\le N}\left[\delta_T^s(j)\right] \quad (7)$$

$$\text{From } s = 1 \text{ to } S \;\Rightarrow\; q_T^s = \operatorname*{argmax}_{1\le j\le N}\left[\delta_T^s(j)\right] \quad (8)$$

where the superscript s implies the association of the s^{th} sub-band, S is the number of sub-bands, N is the number of states in each sub-band model, T is the number of test vectors in each sub-band, $\mathbf{O}_t$ is

the t^{th} test vector, $b_j(\mathbf{O}_t)$ is the probability of observing $\mathbf{O}_t$ in the j^{th} state, $\delta_t(j)$ is the best scores accumulated along a single path, at time *t*, which accounts for $\mathbf{O}_1, \mathbf{O}_2, ..., \mathbf{O}_t$ and ends in i^{th} state, a_{ij} is the probability of transition from state *i* to *j*, $\psi_i(j)$ is the state that maximises $\left(\delta_{t-1}(i) + \log(a_{ij})\right)$, *P* is the score used for verification and *h* is a weighting function which will be defined and discussed in Section 2.1.

Equation (3), (5) and (8) are required only if the following backtracking procedure is to be applied.

From $t = 1$ to S,

From $t = T-1$ to 1 $\Rightarrow$ $q_t^s = \psi_{t+1}^s(q_{t+1}^s)$ (9)

Next *t*

where $\{q_1^s, q_2^s, \cdots\cdots, q_T^s\}$ is the optimum time-alignment path (given in terms of states) for the test vector sequence $\{\mathbf{O}_1^s, \mathbf{O}_2^s, \cdots\cdots, \mathbf{O}_T^s\}$.

From equation (1)−(8) it is evident that, in between two consecutive recombination levels, the operations are the same as those in an approach where the conventional Viterbi algorithm is applied independently in each sub-band at the same time. At recombination levels, however, the δ values which correspond to the same states in different sub-bands are set to their mean values (equation 1 and 6). This allows the time-alignment paths developing in an individual sub-band to get some benefit from the spectral information in the other sub-bands. As a result, the optimum set of time-warping paths and resulting verification score are expected to become more reliable. For the purpose of this paper the above approach is referred to as *sub-band HMM* (*SB-HMM*).

2.1. Estimation of the sub-band weights

An important factor in the modified version of the Viterbi algorithm described above is the choice of *h* weights which operate directly on the band-limited segmental scores. By determining these weights appropriately, it is possible to emphasise the band-limited segments that are more specific to the speaker whilst de-emphasising the ones that are affected by time and/or frequency localised anomalies. The techniques which can be used for this purpose may be grouped into three categories according to the type of the measure involved. These measures are (i) a priori knowledge of the sub-band performance, (ii) the segmental level SNR in each sub-band, and (iii) the average score against a set of competing speaker models. The following gives a description of the approach involved in each of these categories.

i) A Priori Knowledge of the Sub-Band Performance

A method for computing the required *h* weights is based on the use of knowledge of the relative performance of the individual sub-bands. Such knowledge can be gained through a series of experiments using a given set of speech data. For example, if at a given speech unit level (e.g. phoneme, syllable, word), the average verification error rates in the sub-bands 1,2,...,*S* are $(1-\eta_1)$, $(1-\eta_2)$,...,$(1-\eta_S)$ respectively, then the required weights may be specified as

$$h_t^s = \eta_s \Big/ \sum_{s=1}^{S} \eta_s \qquad 1 \le s \le S\,, \quad T' \le t \le T'' \tag{10}$$

where T' and T'' are the boundaries of the considered speech unit. The above formulation indicates that the weights are linearly proportional to the corresponding 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 [17]

$$h_t^s = \mathrm{f}\left(\eta_s / \max(\eta_s)\right) \Big/ \sum_{s=1}^{S} \mathrm{f}\left(\eta_s / \max(\eta_s)\right) \qquad 1 \le s \le S\,, \quad T' \le t \le T'' \tag{11}$$

where $\mathrm{f}(x) = x / \left(1 + 20e^{-6x}\right)$. (12)

Comparing this approach with that described by equation (10) 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 [6][11]. 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, as described below, is to incorporate the localised levels of contamination in the test utterance into the process of generating the weights.

ii) Use of Segmental level SNR in each Sub-Band

In order to reduce the effect of additive band-limited noise, the *h* 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.

$$h_t^s = \frac{1}{S-1}\left\{1-\left(\Phi(s,t)\Big/\sum_{s=1}^{S}\Phi(s,t)\right)\right\}, \tag{13}$$

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 [13]. An example of this is

$$\Phi(s,t) = \frac{1}{K'' - K' + 1}\sum_{k=K'}^{K''}\frac{B_{\max}(k)}{1+\gamma\rho(k,t)}, \tag{14}$$

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 the k^{th} frequency index in the considered noise frames, and $\rho(\cdot)$ is the presumed band-limited frame level SNR which is given as:

$$\rho(k,t) = |X(k,t)| / |B(k,t)|, \tag{15}$$

where $|X(\cdot)|$ and $|B(\cdot)|$ are the estimates for the spectral magnitudes of the smoothed noisy speech and the noise respectively [13].

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 [12]. 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 of 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 described in the next section 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.

iii) Use of Competing Speaker Models

The underlying idea here is to determine the *h* weights dynamically 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 a selected set of background speaker models. Such background models can be either speaker independent sub-band models or a group of sub-band model sets that are capable of competing with the sub-band model set of the target speaker. In the latter case the competing speaker model set can be selected based on their closeness to either the target model set or the test utterance [2]. For the reason stated below, the second approach (which is also unknown as *unconstrained cohort normalisation* [2][4]) was chosen in this study. Based on

this method, the h weights can be defined as follows.

$$\log h_t^s = -\frac{1}{U}\sum_{u=1}^{U} \log b_{x(t)}^{us}(O_{st}), \tag{16}$$

where U is a prefixed value which defines the number of speakers to be allowed in the competing set [2], and $b_{x(t)}^{us}(O_{st})$ is the probability of observing the t^{th} test vector of s^{th} sub-band in the $x(t)^{th}$ state of s^{th} sub-band model belonging to the u^{th} competing speaker. 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 modified Viterbi algorithm described by equations (1) – (8) and then the backtrack procedure described by equation (9) has to be applied (Figure 2). It should be noted that each of these alignment procedures involves an additional set of h weights. These weights can simply be set to 1 or computed according to the band-limited frame level SNRs as described in the previous section (the former approach is used in this study).

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 (this technique is commonly known as bootstrap [16]).

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 token and not necessarily to the target model. As a result h_t^s will become small and thereby the probability of false acceptance will be reduced significantly.

For the purpose of this paper the above method of determining h weights is referred to as dynamic band-limited segmental weights (DBSW). It should be pointed out that the fundamental difference between the DBSW technique and the conventional score normalisation approaches [2] is that the

latter assumes that the mismatch is uniform across the entire frequency range at a given segmental level. The DBSW method, on the other hand, does not make such an assumption and attempts to estimate the level of mismatch associated with each individual band-limited segment of the utterances. This information is then used to compute a weighting factor for correcting each segmental (band limited) distance prior to the calculation of the final score.

3. Sub-band Cepstrum

The main focus of this section is the spectral information represented by the sub-band cepstral parameters in relation to that contained in the full-band cepstral parameters [18]. Suppose that a S sub-band system is formed by equally distributing the log spectral parameters $Y(k)$, $k = 0,1,..,K-1$. The p^{th} cepstral coefficient of the s^{th} sub-band can be computed according to the following equation.

$$c_S(s,p) = \frac{1}{(K/S)} \sum_{k=(s-1)K/S}^{(sK/S)-1} Y(k) f_{p,K/S}(k) \tag{17}$$

where $f_{p,K/S}(k)$ is k^{th} order basis function of the discrete cosine transform (DCT) and it is given by

$$f_{p,K_S}(k) = \cos\left(\frac{a_1 \pi p (k + a_2)}{K_S}\right). \tag{18}$$

The values of a_1 and a_2 embedded in the above cosine basis function are either {2 and 0} or {1 and 0.5} respectively. The former set of values is commonly used in computing the fast Fourier transform (FFT) based cepstral parameters. These values are also implicitly involved in the computation of the LPC based cepstral parameters [8]. In this case, equation (17) implies that

$$c(S \times p) = (1/S) \sum_{s=1}^{S} c_S(s,p) \tag{19}$$

where $c(x)$ { $= c_1(1,x)$ } is the x^{th} full-band cepstral parameter. The latter set of values (i.e. 1 and 0.5) is commonly used in determining mel-frequency cepstral coefficients (MFCCs) [7]. In this case, the relationship between the full and sub band cepstral parameters is of the following form

$$c(S \times p) = (1/S) \sum_{s=1}^{S} (-1)^{p(s-1)} c_S(s,p) \tag{20}$$

From equation (19) and (20), it is evident that there exists a strong linear relationship between $c(S \times p)$ and $\{c_S(1,p), c_S(2,p), ..., c_S(S,p)\}$. In other words, the cepstral coefficients with the same index in different sub-bands have a strong linear relationship to a full-band cepstral parameter whose quefrency is given by the product of that specific index with the number of sub-bands. This would, for example, imply that in a 4-sub-band system, the parameter sets

$\{c_4(1,p), c_4(2,p), c_4(3,p), c_4(4,p)\}_{p=1,2,3,.....}$

have high correlations with the full-band coefficients $c(4)$, $c(8)$, $c(12)$,..... respectively.

In general, all the mid-quefrency full-band cepstral parameters {e.g. $c(3) – c(8)$} are highly useful for speaker discrimination. This is because they are neither, like the lower-order parameters, easily affected by the transmission channels nor, like higher-order parameters, significantly influenced by the voice pitch (which is susceptible to both mimicry and changes over time and also significantly affected by factors such as emotional state and speech efforts [9]). This implies that a linear relationship could not be established between each and every full-band cepstral parameter that is highly useful for speaker discrimination and a set of cepstral coefficients of the same quefrency belonging to an S-sub-band system ($S > 1$). This in turn arises a question: could the final decision in any of the conventional classification methods that are adapted to use only the cepstral parameters derived from an S-sub-band system account for the spectral information contained in all the full-band cepstral parameters that are proven to be useful for speaker discrimination?

The experimental investigations carried out in the HMM framework to determine the answer to this question is described in Section 4. The results of these study shows that indeed the cepstral parameters derived from a single sub-band-system cannot capture all the spectral information contained in the full-band cepstral coefficients that are highly effective for speaker verification. This is obviously a serious obstacle in using the sub-band analysis for the purpose of speaker verification. The following sections provide possible methods to tackle this problem.

Remark I: It should be noted that simple linear relationships, such as (19) and (20), would be possible between sub-band and full-band cepstral parameters only if the log-spectral magnitudes are equally

divided into non-overlapping sub-bands. However, the general point in the above discussions is that the overall spectral variation measured by a lower order DCT basis function in a sub-band is comparable to that of a higher DCT basis function of the full-band in the same spectral region. As a result, the details of the overall spectral variation measured by the complete set of DCT basis functions of the full-band in a given sub-band cannot be extracted using the DCT basis functions associated specifically with that sub-band.

Remark II: Based on the results of a set of tests carried out in the initial part of the experimental investigation (which is presented in Section 4.3), it was decided to choose the frame-level sub-band recombination for the subsequent parts of the experimental study including the evaluation of the methods proposed in the next two sections. Hence, to simplify the discussions in the next two sections, these methods are presented based on the assumption that the sub-band recombination is performed at the frame-level. It should be pointed out that under this condition, the modified Viterbi algorithm described by the equations (1) – (8) reduces to that presented below:

Step 1 : Initialisation **:**

$$\delta_1(1) = S^{-1} \sum_{s=1}^{S} \log\left(b_1^s(O_1^s) h_1^s\right) \tag{21}$$

$$\text{for } j = 2 \text{ to } N,\ \delta_1(j) = -\infty \tag{22}$$

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

$$\delta_t(j) = S^{-1} \sum_{s=1}^{S} \left\{ \max_{1 \le i \le J} \left[\delta_{t-1}(i) + \log a_{ij}^s\right] + \log\left(b_j^s(O_t^s) h_t^s\right) \right\} \tag{23}$$

Step 3 : Termination

$$P = \max_{1 \le j \le N} \left[\delta_T(j)\right] \tag{24}$$

It should also be noted that the operations to determine the optimum set of time-alignment paths were omitted in the above description since they would have the exact same form as before.

3.1. Sub-band system with full-band cepstral supplements

One possible method to tackle the problem stated in the previous section is to supplement the sub-band cepstral coefficients with appropriate subsets of the full-band cepstral parameters. In a *S*-Sub-

band system, this subset may be formed using the full-band cepstral coefficients $c(n)$, where n takes all values except those divisible by S (this is because of the fact that coefficients $c(S)$, $c(2S)$, $c(3S)$, … have strong linear relationships with sub-band cepstral parameters). For example, in the case of a 4-sub-band system, this complementary subset could contain $c(1)-c(3)$, $c(5)-c(7)$ and $c(9)-c(11)$, if the full-band quefrency is truncated to 12. For the purpose of this paper, this method is referred to as "*sub-band system with full-band cepstral supplements* – SSFCS". In order to incorporate this technique into the SB-HMM procedure, the term

$$S^{-1}\sum_{s=1}^{S}\log\left(b_j^s(O_t^s)h_t^s\right)$$

in equations (21) and (23) could be replaced with

$$\alpha S^{-1}\left[\sum_{s=1}^{S}\log\left(b_j^s(O_t^s)h_t^s\right)\right]+(\alpha-1)\log\left(b_j'(O_t')h_t'\right)$$

where α is a combination factor between 0 and 1 (in this study, α was simply set to 0.5), O_t' is the set of t^{th} complementary full-band cepstral coefficients (CFBCCs) of the test data, $b_j'(O_t')$ is the probability for observing O_t' in the j^{th} state of CFBCCs based model of the target speaker and, and h_t' are weighting factors for dealing with speech contamination in a time-localised manner (these weights can be computed dynamically based on an approach similar to DBSW (Section 2.1(iii)) and by using the competing speaker models of CFBCCs). Of course, due to the involvement of the full-band parameters, the benefits of the sub-band processing cannot be fully exploited. It is, however, believed that the full-band parameters can significantly improve the robustness of the target speaker model and thereby a better verification accuracy is achieved.

3.2 Multiple sub-band-systems analysis

Another way of tackling the above problem may be that of ensuring that there is a strong linear relationship between every full-band cepstral coefficient which is proven to be useful for speaker verification, and the sub-band parameters that are to be used in the classification. This can be accomplished if the sets of cepstral parameters are drawn from a group of different sub-band systems. The approach can be further described as follows. Assume that the full-band cepstral coefficients with quefrencies in the range p to $p+q$ are the most useful ones for speaker verification. Based on the

analytical study presented in Section 3, the use of all the 1st cepstral coefficients of $\mathcal{S}(0)$, $\mathcal{S}(1)$,…., and $\mathcal{S}(q)$ (where $\mathcal{S}(i)$ is the sub-band system with $p+i$ bands) would ensure a strong linear relationship between each of $c(p)$, $c(p+1)$,…., and $c(p+q)$, and the sub-band parameters. Such an approach 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 method would be based on using R ($< q$) sub-band systems in which certain sub-band systems contribute more than one set of cepstral coefficients. With this technique, it is possible to avoid the systems that have unusually large number of sub-bands. However, the difficulty is in determining the relationship for the full-band cepstral coefficients $c(i_1)$, $c(i_2)$, $c(i_3)$, . . . , where $p \le i_x \le p+q$ and i_x is not divisible by any of $S_1, S_2, \ldots, S_R$ (S_r is the number of bands in the r^{th} sub-bands system). A possible method to tackle this problem is to supplement $\{c(i_1), c(i_2), c(i_3), \ldots\}$ to the chosen set of sub-band parameters. It should be pointed out that, in this approach, the number of full-band parameters to be used is much smaller than those in the case of SSFCS and thus the benefits of the sub-band analysis can be better realised. For the purpose of this paper, this technique is referred to as *multiple-sub-band-systems analysis* (MSBSA). Figure 3 illustrates a possible choice of parameters in the MSBSA method to represent the full-band cepstral coefficients $c(1)$ – $c(12)$. In order to incorporate this technique in the SB-HMM procedure, the term

$$S^{-1}\sum_{s=1}^{S}\log\left(b_j^s(O_t^s)h_t^s\right)$$

in equations (21) and (23) can be replaced with

$$\alpha R^{-1}\sum_{r=1}^{R}\left[\beta_r S_r^{-1}\sum_{r=1}^{S_r}\log\left(b_j^{rs}(O_t^{rs})h_t^{rs}\right)\right]+(\alpha-1)\log\left(b_j'(O_t')h_t'\right),$$

where $\sum_{r=1}^{R}\beta_r = 1$ {in this study, β_r is simply set to $1/R$}, R is the utilised number of sub-band systems, the superscript r indicates the association of the r^{th} sub-band system, and the symbols α, $b_j'(O_t')$ and h_t' have the same meanings as in the case of SSFCS. This method can certainly provide more flexibility in dealing with time and frequency localised anomalies. Its main drawback, however, is the increase in the computational complexity. It should also be noted that, the number of full-band

cepstral coefficients available to form a speaker model here is far less than that used in the conventional full-band analysis. Furthermore, to build speaker models for an associated sub-band system, it may not be possible to use the entire set of cepstral parameters that would have been chosen in an analysis based solely on that sub-band system. This can lead to an unreliable speaker representation in the HMM framework. In order to tackle this problem a collective training method is devised and applied. The details of this technique are presented in the appendix.

4. Experimental Investigation and Results

The experimental work is divided into two main parts. In the first part, a number of issues concerning the techniques discussed in Section 2 and 3 are addressed. This includes the usefulness of different spectral regions for speaker verification as well as the effects of increasing the number of sub-bands and relaxing the time-synchrony assumption. The details of these studies are covered in Sections 4.2 and 4.3. In the second part, the robustness of the proposed methods under adverse conditions is evaluated. The details of this part of the investigation are presented in Sections 4.4 and 4.5. The purpose of Section 4.1 is to present the general considerations in the experimental studies.

4.1. General considerations

i) Definition of sub-bands

A critical choice to be made in the design of a sub-band based system was the number and the positions of the constituent sub-bands. Obviously, narrower sub-bands would allow greater flexibility in dealing with the frequency localised anomalies. However, in such cases, the sub-bands would contain less information regarding the speaker, which could lead to unreliable partial decisions. The sub-band systems used for the purpose of this study were chosen in relation to the critical filterbank configuration used in the feature extraction procedure described earlier. In each of these systems, the individual bands were set such that they covered equal number of critical bands while the overlap between the consecutive bands were kept as minimum as possible (more precisely, no more than one overlap between consecutive sub-bands was allowed to fully cover a critical band in the corresponding frequency region). Examples of the sub-band systems used in this work are given in Figure 4.

ii) Speech data

The speech data chosen for this study was a subset of the *BT Millar speech database*. This database was collected in a quiet environment over a period of approximately three months using a high quality microphone. It consists of 25 repetitions of digit utterances 1 to 9, "zero", "oh" and "nought" spoken by a total of 63 native English-speakers. Each speaker participated in five sessions and repeated the above utterances five times in each session. Subjects were prompted to speak individual digit utterances in a random order. All utterances were validated by human listeners, and provided with approximate endpoints within the digital recording. The first 10 versions of each utterance (obtained over the first two recording sessions) were reserved for training and the last 15 versions of the utterances (recorded over the last three sessions) formed the standard test set of the database.

The subset adopted for the purpose of this study consisted of digit utterances 1 to 9 and "zero" spoken by 20 speakers. This subset has a bandwidth of 3.1 kHz (telephone bandwidth) and a sample rate of 8.0 kHz. In order to make the speaker verification task more challenging, all the speakers in this set were chosen to be male speakers of about the same age.

iii) Feature parameters

The main type of feature parameters chosen for the purpose of this study was SB-MFCC. 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 8^{th} order FFT. The critical filterbank configuration used in this study was of the type proposed in [7]. An additional aim of the experimental study was to investigate how the spectral information contained in SB-MFCCs was influenced by DCT (Section 3). It was therefore decided to provide a basis for this investigation by producing a second type of feature parameters using only the log-energy outputs of the filterbank in each sub-band. For the purpose of this paper, these alternative features are abbreviated as SB-MFBOs (sub-band, mel-scale filterbank outputs).

The full-band feature sets that were used for the purpose of comparison were MFBOs and MFCCs. The former is the entire log-energy outputs of the filterbank and the latter is the result of applying the DCT to the corresponding set of MFBOs.

iv) Speaker representation

In all the experimental investigations discussed in the subsequent sections, the speaker representation was based on digit-level HMMs. The HMM topology adopted was 4s2m, where *N*s*M*m stands for *N*-state left-to-right structure without any "skip" transitions and *M* Gaussian mixtures per state (it should be noted that this selection of topology was made on the basis of a set of preliminary study carried out to evaluate the relative effectiveness of 2s4m, 4s2m, 2s6m and 8s1m). The *k*-means algorithm was used to train the speaker models [15]. In this procedure, for a given utterance text, the same speaker-independent-HMM was used as the seed in the training of all the associated speaker models. Furthermore, when cepstral feature parameters were used, the covariance matrixes of the probability distributions associated with the speaker models were assumed to be diagonal.

v) Testing procedure

In each verification trail, each of the reference models of each registered speaker for each vocabulary item was compared against the corresponding test set of all 20 speakers. Hence, the total number of true speaker and impostor tests carried out in a verification trail were $\mathcal{T}_1 = n_t \times n_r \times n_v$ and $\mathcal{T}_0 = n_t \times n_i \times n_r \times n_v$ respectively. Here, n_t (=15), n_r (=20), n_v and n_i (=19) have the following respective meanings: the number of test versions, the number of registered speakers, the number of vocabulary items, and the number impostors tested against each given registered speaker. The experimental investigation given in Section 4.5(iii) was based on a vocabulary of three sequences of four-digits which implies that, in this case, $\mathcal{T}_1 = 900$ and $\mathcal{T}_0 = 17100$. In all the other experimental studies, the vocabulary items were the 10 individual digits in the set resulting in $\mathcal{T}_1 = 3000$ and $\mathcal{T}_0 = 57000$. The final scores obtained in each verification trail were used to form the empirical distributions of the true speakers and impostors to determine the equal error rate (EER) i.e., the probability of equal number of false acceptances and false rejections.

The above discussion indicates that, in this study, the false rejection rate involved in the EER computation is subject to much larger statistical variation than the corresponding false acceptance rate, as it is based on relatively fewer trials. In this case, an estimate of the 95% confidence interval (CI_{95}) for EERs may be obtained by setting the number of constrained trials equal to the number of true speaker tests, and using a single normal distribution. Based on this, an estimate of the 95% confidence interval is given by

$$CI_{95} = \varepsilon \pm 1.96\sqrt{\varepsilon(100-\varepsilon)/\mathcal{T}_1}\,, \tag{25}$$

where ε is the EER in percentage and $\mathcal{T}_1$ has the same meaning as that given earlier. For example, in the case of single digit utterances, the 95% confidant intervals for 5% and 15% EERs are (5 ± 0.8)% and (15 ± 1.3)% respectively.

4.2. Relative effectiveness of different sub-bands

In this study, the experiments were carried out separately for every band in each considered sub-band system {Section 4.1(i)} as well as for the full-band system using the conventional HMM framework. The feature parameters used were SB-/MFBOs, and SB-/MFCCs {Section 4.1(iii)}. The purpose of this investigation was twofold. Firstly, to acquire the knowledge of how relevant the different spectral regions for speaker verification were. Of course, such knowledge is crucial if one chooses to use weighting scheme described in Section 2.1(i). Secondly, to understand how the effectiveness of speaker verification is influenced by the use of an independent DCT in each sub-bands.

The results of this investigation are given in Figure 5 as a function of the sub-band frequencies. Based on these results it can be said that, in general, the mid- and high-frequency sub-bands (1000-2500 & 2500-4000 Hz) are more useful for speaker verification than the low-frequency ones (100-1000 Hz). In particular, the mid-frequency sub-bands appear to have the highest ability to verify speakers. It is also interesting to note that in the case of 2-sub-band SB-MFBOs, the band encompassing the upper part of the spectral region achieves slightly better performance than that of the conventional full-band.

Remark*:* In reviewing the speaker recognition literature, the authors have come across two studies on the relative performance of sub-bands in a system covering the entire frequency band of interest [3][5]. Both of these studies, which contradict the above results, are concerned with the closed-set speaker identification tasks. In the first study, the experiments have been conducted in a text-independent mode and it has been found that the sub-bands in region 600 – 2000 Hz are the most irrelevant ones for the task undertaken. In the latter study, the experiments have been carried-out in a text-dependent manner using a digit database. The results have shown that if all the speakers are male, the sub-bands in the regions 100 – 500 Hz and 1200 - 2600 Hz are relatively better for speaker identification and the identification error is at its peak for the sub-bands in the region 700 – 900 Hz. In this study, it has also been found that the relative performance of sub-bands for female speakers is somewhat different from that of the male speakers. These results together with that reported here suggest that the speaker discrimination effectiveness of some parts of the frequency region largely

depends not only on the chosen phonetic content and the type of the speaker group (particularly the gender) but also on the type of speaker recognition.

Figure 5 also shows a general increase in the verification error rate for both groups of feature parameters as the number of sub-bands is increased. This increase is observed to be relatively large in the case of SB-MFCCs. The dependence of the verification error rate on the number of sub-bands can be further understood by viewing Figure 6. The plots in this figure are based on the EERs computed using the values of P_S, where $P_S = S^{-1}\sum_{i=1}^{S} P_S(i)$ and $P_S(i)$ is the final score obtained independently in the i^{th} band of the sub-band-system which consists of S bands. The obvious and the common reason for the observed increase in the EER for both types of feature parameters is the reduction of the spectral information as the sub-band is narrowed. The cause for the additional increase in the case of SB-MFCCs has to be attributed to the way in which these parameters represent the spectrum (Section 3).

4.3. Comparative examination of various recombination levels

Based on the discussion provided in Section 2, it is evident that the level in which sub-bands are recombined is an important factor in using the SB-HMM method effectively. Therefore, a series of experiments were carried out to determine the best recombination level. In this study the number of sub-bands was arbitrarily set to four. The recombination levels examined in the initial stage of the investigation were single frame, phoneme and word. In order to obtain the required phonetic boundaries in the second case, each test utterance was forced to align against phoneme-based (speaker-independent) full-band HMM of the corresponding digit. Table 1 presents the results of this study in terms of EER for SB-MFBOs. According to these results, single frame and word are the most and the least effective recombination levels respectively. A set of experimental study conducted using SB-MFCCs also led to similar results.

A possible reason for the above order of effectiveness and for not realising the benefits of relaxing the time-synchrony assumption is that the duration between the consecutive time-resynchrony points is

too long. In such case, due to the extensive use of partial information, the time-warping paths become less reliable. It was thought that this problem could, to a certain extent, be overcome by selecting a merging level slightly higher than a single frame. Of course such levels might not necessarily be the ideal time-resynchrony points. However, they would provide a certain degree of freedom in choosing the time-warping paths of different sub-bands.

In order to investigate this idea experimentally, a set of verification tests were carried out by incrementing the merging level from 1 to 10 frames. The results of this study are summarised in Figure 7. These results confirm that a better verification accuracy can be achieved by using the suggested approach. However, the merging level which yielded the best performance varied considerably depending on the utilised utterance text (see the tabulation at the bottom of figure). Because of this, it was decided to use the single-frame-level merging for the remaining parts of the experimental work to keep the verification algorithm as simple and efficient as possible.

4.4. Experimental investigations with the DBSW technique

For this part of the experimental investigation, the effect of the time-localised anomalies was simulated by adding an arbitrarily chosen narrow band noise (0-600 Hz) to randomly set time-regions of each test utterance. The net duration of contamination in each test utterance was kept close to 1/3 of its length. Moreover, the signal to noise ratio (SNR) was maintained at a pre-determined level. The number of sub-bands used was four.

i) Baseline experiments

The first set of experiments was aimed to investigate the effectiveness of the sub-band HMM (SB-HMM) approach in the log spectral domain. For this purpose, SB-MFBOs were chosen and the DBSWs were applied (Section 2.1(iii)). The results of this study are given in Figure 8. In order to perform a meaningful comparison, the figure also includes the results obtained under similar experimental conditions using three other techniques. These techniques are the standard full-band HMM

(FB-HMM), FB-HMM with unconstrained cohort normalisation (FB-HMM + UCN) and, SB-HMM with SNR-based *h* weights (SB-HMM + SNRW). In the case of FB-HMM + UCN, the cohort size was set to 2. This choice was based on the empirical evidence provided in [2]. Moreover, the SNR-based weights used in the case of SB-HMM + SNRW were determined according to equation (13). This is because the utterances adopted in the experiments were relatively shorter in duration and the noise used to contaminate them was reasonably stable.

The robustness of the SB-HMM + DBSW method is clearly evident from these result (it can be seen that the response of this method across the considered SNR is relatively flat). The figure also shows that by incorporating the UCN in the FB-HMM method the effect of time and frequency anomalies can be compensated to some extent. Furthermore, the SB-HMM + SNRW technique appears to work reasonably well in normalising the effect of applied interference. This may be expected, because the contamination here is due to additive noise. It should, however, be emphasised that this method, unlike the SB-HMM + DBSW and FB-HMM + UCN techniques, cannot be very useful for minimising the effects of such causes of mismatch as the speaker generated variation.

ii) Experiments with SB-/MFCCs

In the next part of the investigation, the above experiments were repeated using SB-MFCCs. The results of this study are given in Figure 9. A comparison of these results with those in Figure 8 indicates that, for both full and sub-band systems, MFCCs perform better than MFBOs. It is also observed that the SB-HMM + DRW method again exhibits a relatively flat response across the considered SNR range. However, the overall performance of the FB-HMM + UCN is noticeably better than that of the SB-HMM + DRW approach. Based on this it can be concluded that in the considered sub-band based speaker verification procedure, the sub-band cepstral parameters do not capture all the spectral information contained in the full-band cepstral parameters that are effective for speaker verification.

Figure 9 also shows the results obtained using the SSFCS and MSBSA techniques (Sections 3.1 and 3.2), which were developed to tackle this particular deficiency of the sub-band cepstrum. In the case

of SSFCS, a 4-sub-band system was chosen whereas in the case of MSBSA, sub-band-systems 2 – 4 were used. These results clearly indicate that MSBSA is more effective than all the considered approaches. It should be, however, noted that the SSFCS technique, which achieves noticeably better performance than the FB-HMM+UCN method, is computationally less expensive than the MSBSA approach.

4.5. Experimental studies using various types of noise

The experimental results presented in the previous section clearly indicated that when the test utterances were partially contaminated by one type of additive narrow band noise, the performance of MSBSA was superior to all other methods considered. This section details the experimental investigations conducted using various types of additive noise in order to scrutinise the effectiveness of the MSBSA technique further. In this study, the results obtained for the FB-HMM+UCN were assumed to be the baseline. Furthermore, the test utterances were contaminated over their entire duration by adding a selected type of noise to them in the time domain at a predetermined SNR of x dB (where x is 10 unless specified otherwise).

i) Experiments using narrow band noise types

The first set of experiments was aimed to investigate the effect of speech contamination in different frequency bands on the considered speaker verification methods. For the purpose of this study, the full frequency range was divided into four regions, $\mathcal{A}_1 \equiv$ (0 – 500 Hz), $\mathcal{A}_2 \equiv$ (500 – 1000 Hz), $\mathcal{A}_3 \equiv$ (1000 – 2000 Hz) and $\mathcal{A}_4 \equiv$ (2000 – 4000 Hz), which were equally spaced in the mel-scale. Each of these regions was in turn contaminated more heavily than the rest with filtered white noise and in each case a set of verification tests was conducted. Figure 10 shows the power spectra of the utilised types of noise and how they influence the average SNR in the frequency domain. The speaker verification error rates for the considered four types of noise are shown in Figure 11.

These results show that a heavy contamination of the signal in the high-frequency regions degrade the verification accuracy much more than that resulting from a heavy contamination of it in the low-frequency regions (it is seen that the EER for the case of $\mathcal{A}_4$ is about 5.5% higher than that for the case of $\mathcal{A}_1$). This must be attributed to the fact that the high-frequency parts of the speech spectrum, which are inherently low in energy, are more effective for speaker verification than the low-frequency parts. These results also show that the considered MSBSA-based method is more effective when the low-frequency regions are heavily contaminated. In this case, the reduction achieved in EER is about 30% of the corresponding baseline value. With this method, the lowest reduction in the EER (about 15% of the baseline value) is obtained when the $\mathcal{A}_4$ region is heavily contaminated.

ii) Experiments using white and pink noise types

In the next set of experiments, two commonly known types of noise, namely *white* and *pink*, were used in turn to contaminate the speech utterances. Figure 12 shows the power spectra of these noise types and their effects on the average SNR in the frequency domain. The results of the verification tests carried out in this part of the experimental investigation are presented in Figure 13.

As expected, when the additive noise is white, the verification accuracies are lower than those in the case of pink noise. It is, however, interesting to note that, even with white noise, the considered MSBSA-based method achieves some reduction in the verification error rate. In this case, the reduction is about 4% of the baseline EER. This may be due to the fact that the low-frequency parts of the speech spectrum, which inherently contain more energy than the high-frequency parts, are less affected by white noise. As seen in Figure 12, the pink noise contaminates the region (0 – 600 Hz) more heavily than the rest of the spectrum. It is also observed that in the said region, pink noise energy is higher than that of white noise. On the other hand, in the high-frequency regions (> 2 kHz), pink noise level is about 10 dB less than that of white noise. Figure 13 shows that the EER for pink noise is about 10% less than that for white noise and the reduction in EER due to the use of MSBSA is about 8% of the baseline.

iii) Experiments using real noise types

For the final part of the experimental investigation, the BT Piper database was chosen. This database contains various noise types, each recorded in a real-life situation using either a land or a cellular telephone. In order to provide a general coverage of this database the following noise types were used in the verification tests.

1. Varied crowd noise in a shopping centre.
2. Constant light chats and keyboard clicks in an office.
3. Car travelling at 30 mph on a busy road with varied surface condition.
4. Car travelling at 70 mph on a busy road with a constant surface condition.
5. Crowd talk and occasional clatter in a restaurant.
6. Background music.
7. High-level variable babble in a pub.
8. Pay phone at roadside
9. Airport lounge with some light crowd noise and foot steps.
10. Voices from TV and radio in the background

In this study, prior to contaminating the test utterances, each noise type was scaled whilst maintaining its relative mean level in the noise data set. Here, unlike in the experiments reported so far, sequences of four digits were used in the verification tests. It was believed that this would simulate the practical situations more closely. In order to reduce the experimental load, only three four-digit-sequences were used. These sequences were determined in the following manner. Firstly, two mutually exclusive digit-sets were selected arbitrarily to form two unique four-digit sequences. The remaining two digits and one chosen arbitrarily from each of the first two sequences were then used to form the third sequence. In this way, the full digit-set was invoked in the experiments. The required speaker models for each of these sequences were obtained through a straightforward concatenation of the corresponding digit-level HMMs that were generated and utilised in the previous experimental studies. Figure 14 shows the power spectra of the adopted types of real noise and how they influence the average SNR in the frequency domain. Figure 15 gives the average of the EERs obtained for the three four-digit-sequences as a function of the considered real noise types.

These results indicate that for all the considered real noise types, MSBSA performs better than the approach based on a combination of full-band HMM and UCN. As before, it is seen that MSBSA is more effective when the high-frequency parts of the spectra are less contaminated than the low-frequency ones, and is less effective when the noise level is relatively flat over all the frequencies of interest.

5. Conclusion

This paper has been concerned with various issues related to the sub-band based text-dependent speaker verification. It has been shown how the sub-band analysis can be incorporated into the conventional speech feature extraction process to determine sub-band cepstral parameters. The investigations have indicated that for a sub-band system of S bands, the cepstral coefficients with the quefrency of p have a strong linear relationship to the $(S\times p)^{th}$ full-band cepstral parameter. It has been shown experimentally that this means the conventional classification methods adapted to work with sub-band cepstral parameters may not be able to capture all the useful spectral information contained in the full-band cepstral parameters. To tackle this problem, two methods have been described. The first approach has been based on supplementing the sub-band cepstral coefficients with appropriate subsets of the full-band cepstral parameters. The second technique has involved the use of cepstral parameters generated from different sets of sub-band systems. Based on the experimental results it has been shown that the latter method is considerably more effective. The main drawback of this technique, however, is that it increases the computational complexity.

The problem of incorporating the sub-band analysis into the conventional classification methods for text-dependent speaker verification has also been addressed. It has been indicated that an important issue in this context is the determination of a set of weights for emphasising the band-limited segments that are specific to the target speaker while de-emphasising or removing the contaminated ones. Three classes of methods for determining the required weights have been proposed and investigated. It has been demonstrated experimentally that out of these, only the methods based on

using competing speaker models are capable of introducing robustness against different types of time and/or frequency-localised anomalies in speech.

The results of the experimental study have shown that the mid- and high-frequency sub-bands (1000-2500 & 2500-4000 Hz) are more speaker specific than the low-frequency ones (100-1000 Hz). These results have also indicated that the use of a sub-band recombination level which is slightly higher than a single frame leads to a higher verification accuracy than that obtainable with any of the standard merging levels (i.e. frame, phone, word).

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] Auckenthaler R., Carey M. and Lloyd-Thomas H., "Score normalization for text-independent speaker verification system," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.

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

[6] 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.

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

[8] Deller, J. R. Jr, Proakis J. G. and Hansen J. H. L, *Discrete-Time Processing of Speech Signals*, Macmillan Inc., New York, 1993.

[9] Doddington G. R., "Speaker recognition-identifying people by their voices," *Proceedings of the IEEE*, Vol. 73, pp. 1651-1664, 1985.

[10] Hayakawa S. and Itakura F., "Text-dependent speaker recognition using the information in the higher frequency band," *Proc. ICASSP'94*, pp. 137-140.

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

[12] 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.

[13] 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 Communication*, vol. 11, pp. 215-228, 1992.

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

[15] Rabiner L. R., "A Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE,* vol. 77, no. 2, pp. 257-285, Feb. 1989.

[16] Rosenberg A. E., Lee C-H., Gokcen S., "Connected word talker verification using whole word hidden Markov models," *Proc. ICASSP'1991*, pp. 381-384, 1991.

[17] 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.

[18] Sivakumaran P. and Ariyaeeinia A. M., "The use of sub-band cepstrum in speaker verification," *Proc. ICASSP'2000*, Vol. II, pp. 1073-1076, June 2000.

[19] Sivakumaran P. and Ariyaeeinia A. M., "Multiple sub-band systems for speaker verification," *Proc. ICSLP'2000*, Vol. 2, pp. 458-461, Oct. 2000.

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

Appendix

In order to conduct the MSBSA in the HMM framework, each registered speaker is represented by using $\left\{\boldsymbol{\lambda}^{rs}\right\}_{\substack{s=1,2,\ldots,R \\ r=1,2,\ldots,S_r}}$, where $\boldsymbol{\lambda}^{rs}$ is the N-state, M-mixture, left-to-right HMM associated with the s^{th} band of the r^{th} sub-band-system which consists of S_r bands, and $S_1 = 1$ (i.e. $r = 1$ corresponds to the full-band). The model parameters associated with each of these speaker representations are estimated using a set of L training utterances, $\{\mathbf{O}^l\}_{l=1,2,\ldots,L}$, where $\left\{\mathbf{O}_t^l = \{\mathbf{o}_t^{lrs}\}_{\substack{r=1,2,\ldots,R \\ s=1,2,\ldots,S_r}}\right\}_{t=1,2,\ldots,T}$ and $\mathbf{o}_t^{lrs}$ is the set of cepstral coefficients chosen from the s^{th} band in the r^{th} sub-band system of the l^{th} training utterance. Based on the discussion provided in Section 3.2, it is clear that for all (or at least some) values of r, $\mathbf{o}_t^{lrs}$ does not include all the cepstral parameters that would have been utilised in an analysis based solely on the r^{th} sub-band system. This can lead to unreliable estimates for the model parameters. In order to tackle this problem a modified version of the Baum-Welch re-estimation procedure is used in this work [19]. In this approach, given the training utterance $\mathbf{O}^l$, the probabilities $\xi_t^l(i,j)$ and $\gamma_t^l(j,m)$ are assumed to be equal for all the HMMs used to represent the target speaker and are computed collectively {$\xi_t^l(i,j)$ is the probability of being in state i at time t, and state j at time t+1, and $\gamma_t^l(j,m)$ is the probability of being in state j using m^{th} mixture component}.

Suppose the parameters of the HMM in the s^{th} band of the r^{th} sub-band system are denoted as follows:

$$\left\{\mathbf{A}^{rs} = [a_{ij}^{rs}]_{N\times N},\ \mathbf{C}^{rs} = [c_{jm}^{rs}]_{N\times M},\ \boldsymbol{\mu}^{rs} = [\boldsymbol{\mu}_{jm}^{rs}]_{N\times M},\ \mathbf{U}^{rs} = [\mathbf{U}_{jm}^{rs}]_{N\times M}\right\}$$

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

$$a_{ij}^{rs} = a_{ij} = \sum_{l=1}^{L}\sum_{t=1}^{T_l-1}\xi_t^l(i,j) \Big/ \sum_{l=1}^{L}\sum_{t=1}^{T_l-1}\sum_{j=1}^{N}\xi_t^l(i,j) \tag{A.1}$$

$$c_{jm}^{rs} = c_{jm} = \sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(j,m) \Big/ \sum_{l=1}^{L}\sum_{t=1}^{T_l}\sum_{m=1}^{M}\gamma_t^l(j,m) \tag{A.1}$$

$$\boldsymbol{\mu}_{jm}^{rs} = \sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(j,m)\cdot\mathbf{o}_t^{lrs} \Big/ \sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(j,m) \tag{A.2}$$

$$\mathbf{U}_{jm}^{rs} = \frac{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(j,m)\cdot(\mathbf{o}_t^{lrs}-\boldsymbol{\mu}_{jm}^{rs})(\mathbf{o}_t^{lrs}-\boldsymbol{\mu}_{jm}^{rs})'}{\sum_{l=1}^{L}\sum_{t=1}^{T_l}\gamma_t^l(j,m)} \tag{A.3}$$

where

$$\xi_t^l(i,j) = \alpha_t^l(i)a_{ij}b_j(\mathbf{O}_{t+1}^l)\beta_{t+1}^l(j)\big/P_l\,, \tag{A.4}$$

$$\gamma_t^l(j,m) = \left(\frac{P_l}{\sum_{l=1}^{L}P_l}\right)\left(\frac{\sum_{i=1}^{N}\xi_t^l(j,i)}{\sum_{j=1}^{N}\sum_{i=1}^{N}\xi_t^l(j,i)}\right)\left(\frac{c_{jm}\mathcal{b}_{jm}(\mathbf{O}_t^l)}{b_j(\mathbf{O}_t^l)}\right) \text{ and} \tag{A.5}$$

$'$ denotes the transpose operation.

In equation (A.4), the values $\alpha_t^l(i)$ and $\beta_{t+1}^l(j)$ 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.

$$\alpha_{t+1}^l(j) = \left(\sum_{i=1}^{N}\alpha_t^l(i)a_{ij}\right)b_j(\mathbf{O}_{t+1}^l), \quad \begin{matrix} t = 1,2,...,T-1 \\ j = 1,2,...,N \end{matrix} \tag{A.6}$$

with the initial conditions

$$\alpha_1^l(1) = b_1(\mathbf{O}_1^l) \text{ and } \alpha_1^l(i) = 0\,,\ i = 2,3,...,N, \text{ and} \tag{A.7}$$

$$\beta_t^l(i) = \sum_{j=1}^{N}a_{ij}b_j(\mathbf{O}_{t+1}^l)\beta_{t+1}^l(j) \qquad \begin{matrix} t = T-1, T-2,...,1 \\ i = 1,2,...,N \end{matrix} \tag{A.8}$$

with the initial conditions

$$\beta_T^l(1) = 1 \text{ and } \beta_T^l(i) = 0,\ \ i = 2,3,...,N. \tag{A.9}$$

The computation of both forward and backward probabilities involve a large number of multiplications of numbers less than unity which, in practice, can cause arithmetic underflow conditions. In order to prevent this, the scaling procedure described in [15] can be adopted. In this case, the probability of generating $\mathbf{O}^l$ by the model set $\left\{\boldsymbol{\lambda}^{rs}\right\}_{\substack{r=1,2,....,R \\ s=1,2,....,S_r}}$, P_l, (which is used in equation A.5) can be determined by using the associated scaling factors [15].

In the above equations, $b_j(\mathbf{O}_t^l)$ represents $\sum_{j=1}^{M} c_{jm} b_{jm}(\mathbf{O}_t^l)$ and the probability $b_{jm}(\mathbf{O}_t^l)$, which first appeared in equation (A.5), is estimated in the following manner

$$\log b_{jm}(\mathbf{O}_t^l) = \frac{1}{R} \sum_{r=1}^{R} \left(\frac{g_t^r}{S_r} \sum_{s=1}^{S_r} \log\left(h_t^{rs} \mathcal{N}_p(\mathbf{o}_t^{lrs}, \boldsymbol{\mu}_{jm}^{rs}, \mathbf{U}_{jm}^{rs}) \right) \right) \tag{A.10}$$

where $\mathcal{N}_p$ is a p-dimensional Gaussian density function with the mean vector $\boldsymbol{\mu}_{jm}^{rs}$ and the covariance matrix $\mathbf{U}_{jm}^{rs}$, h_t^{rs} and g_t^r are weights that control the contribution of the different sub-band systems and their bands respectively (for the purpose of the study presented in this paper both h_t^{rs} and g_t^r were set to 1).

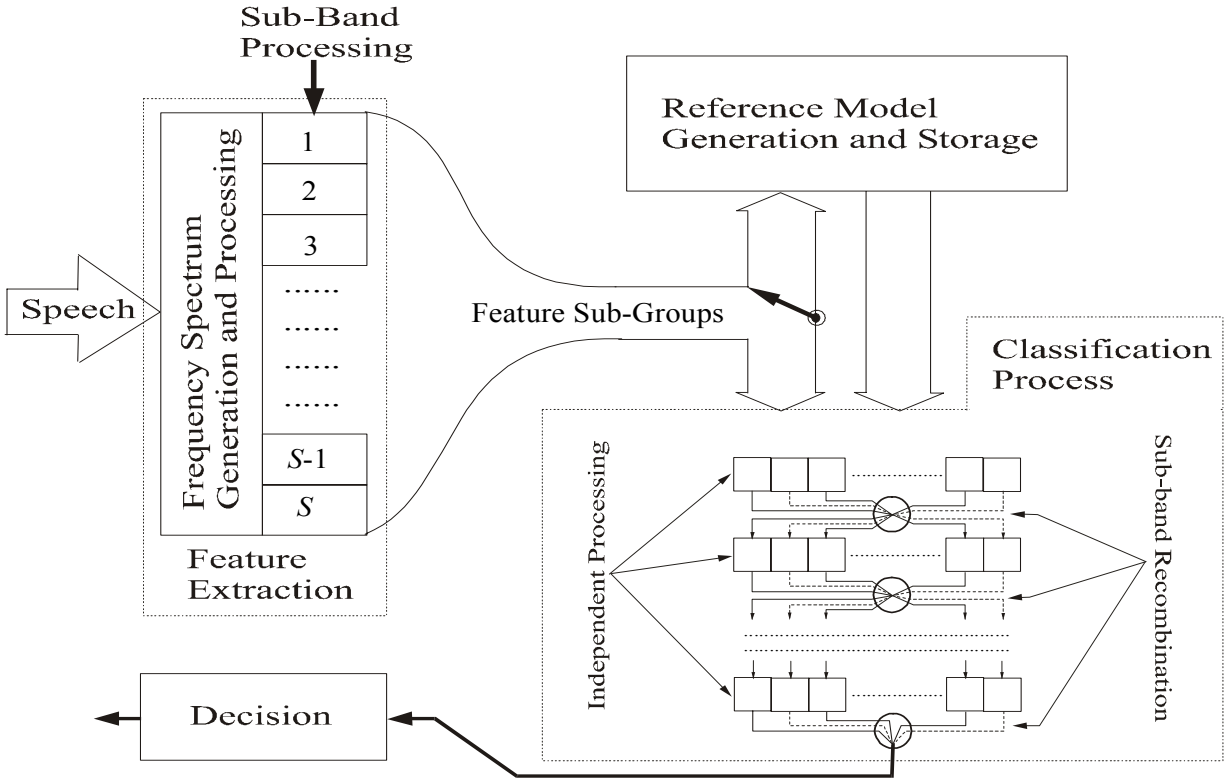


Figure 1: Sub-band based speaker verification system (*S*: number of sub-bands).

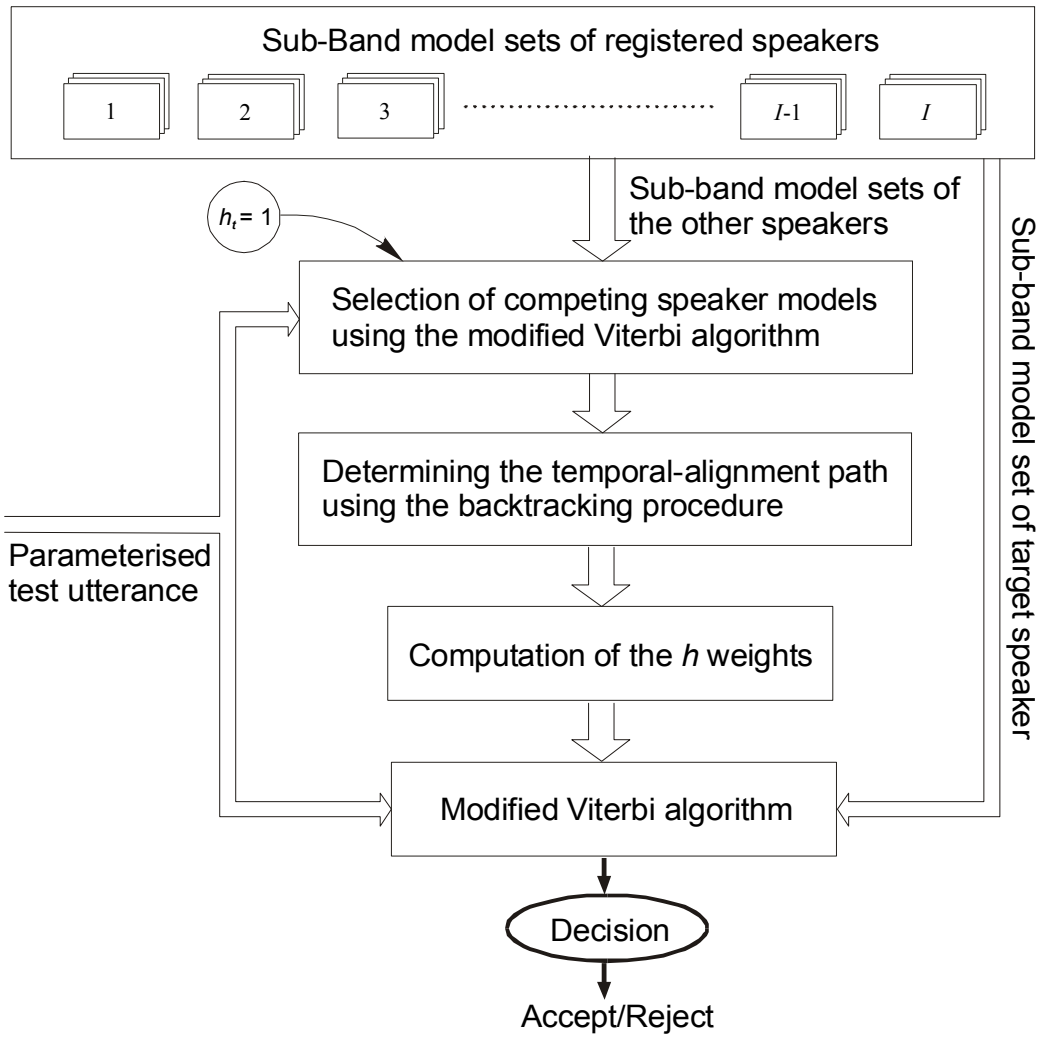


Figure 2: SB-HMM based speaker verification system that uses the sub-band model sets of the competing speakers to determine the *h* weights of the target speaker (*I*: number of registered speakers).

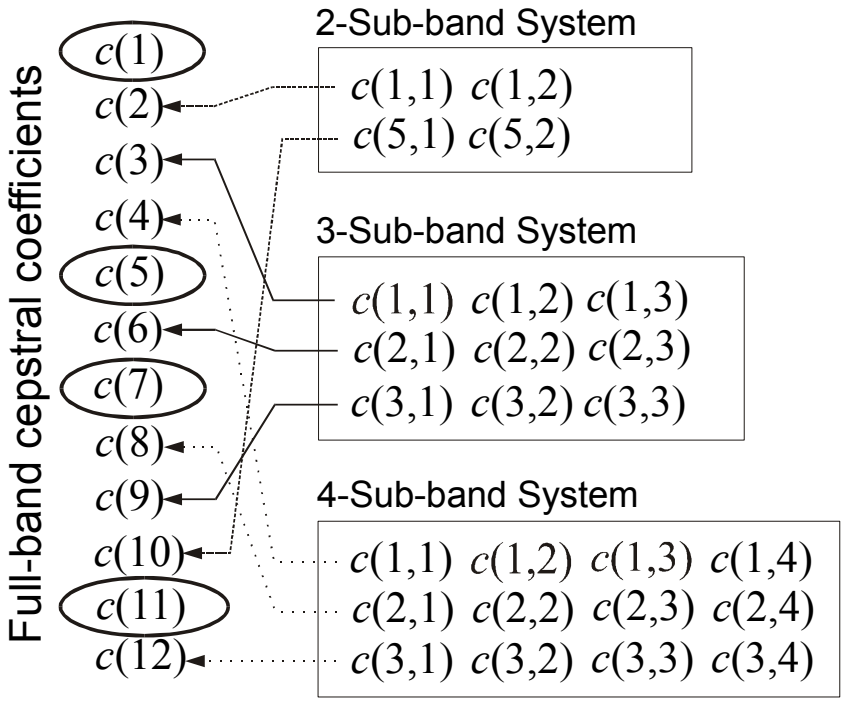


Figure 3: Possible representation of the full-band cepstral sequence with a quefrency span of 1 – 12 in the MSBSA method (the circled full-band coefficients are used as supplementary parameters).

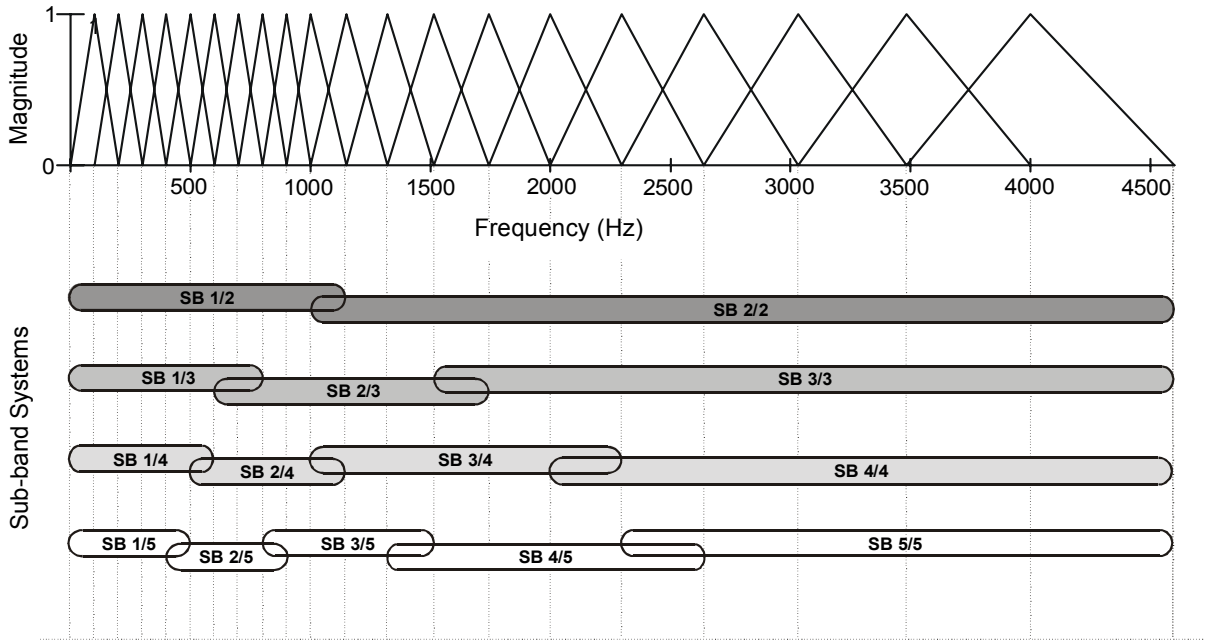


Figure 4: Four different sub-band systems are shown relative to the adopted critical filterbank configuration (SB n/N implies n^{th} frequency band of an N-sub-band system).

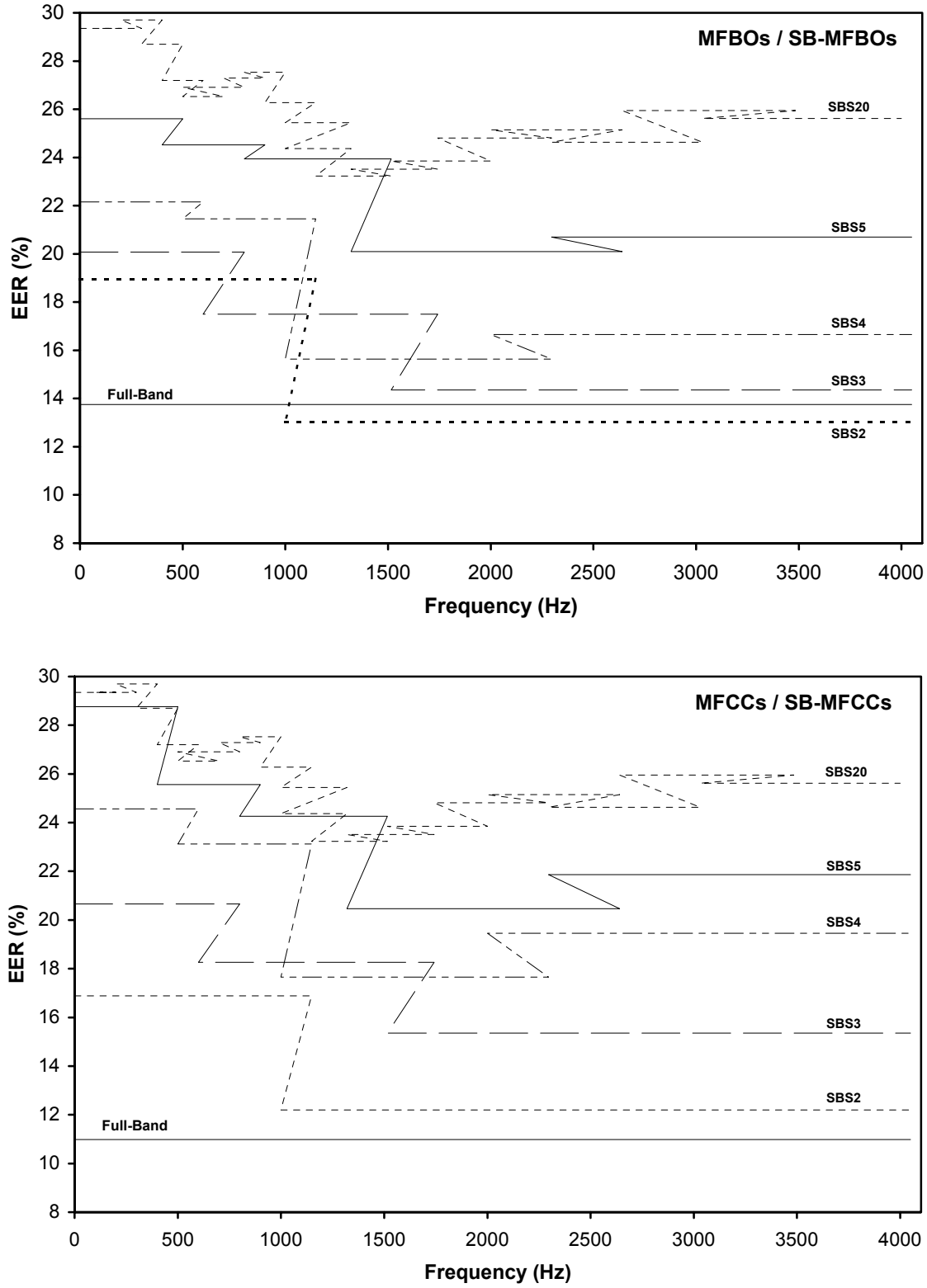


Figure 5: EER as a function of frequency for various sub-band systems (SBS*N* stands for sub-band system of *N* bands. The frequency span of each band is represented as a flat line).

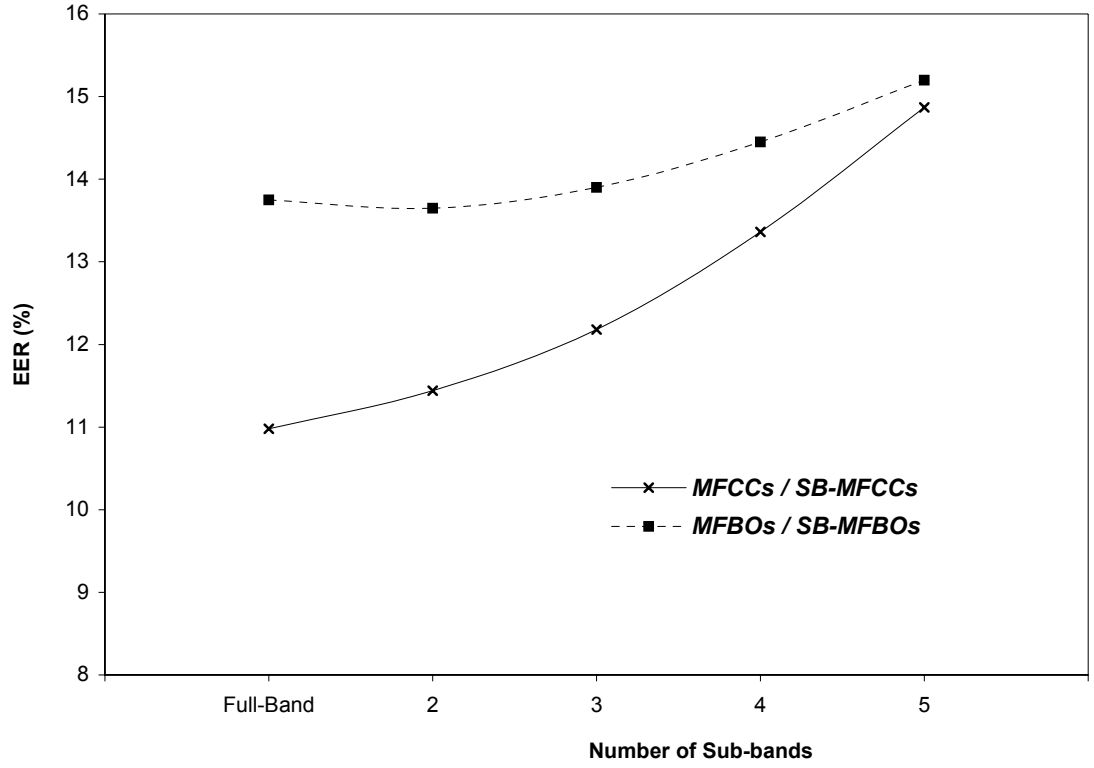


Figure 6: Effects of increasing the number of sub-bands on speaker verification for two different cases.

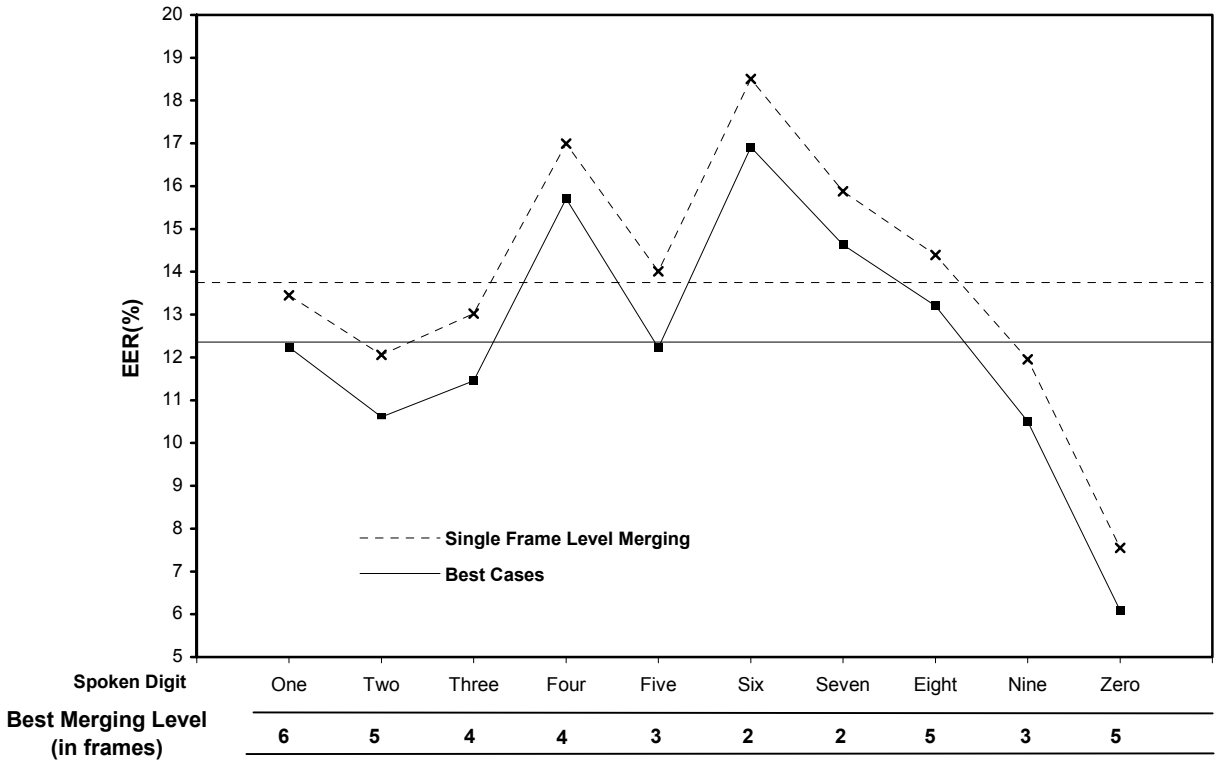


Figure 7: Dependence of the verification error rate on the spoken material and the recombination level (the flat lines represent the corresponding overall EERs).

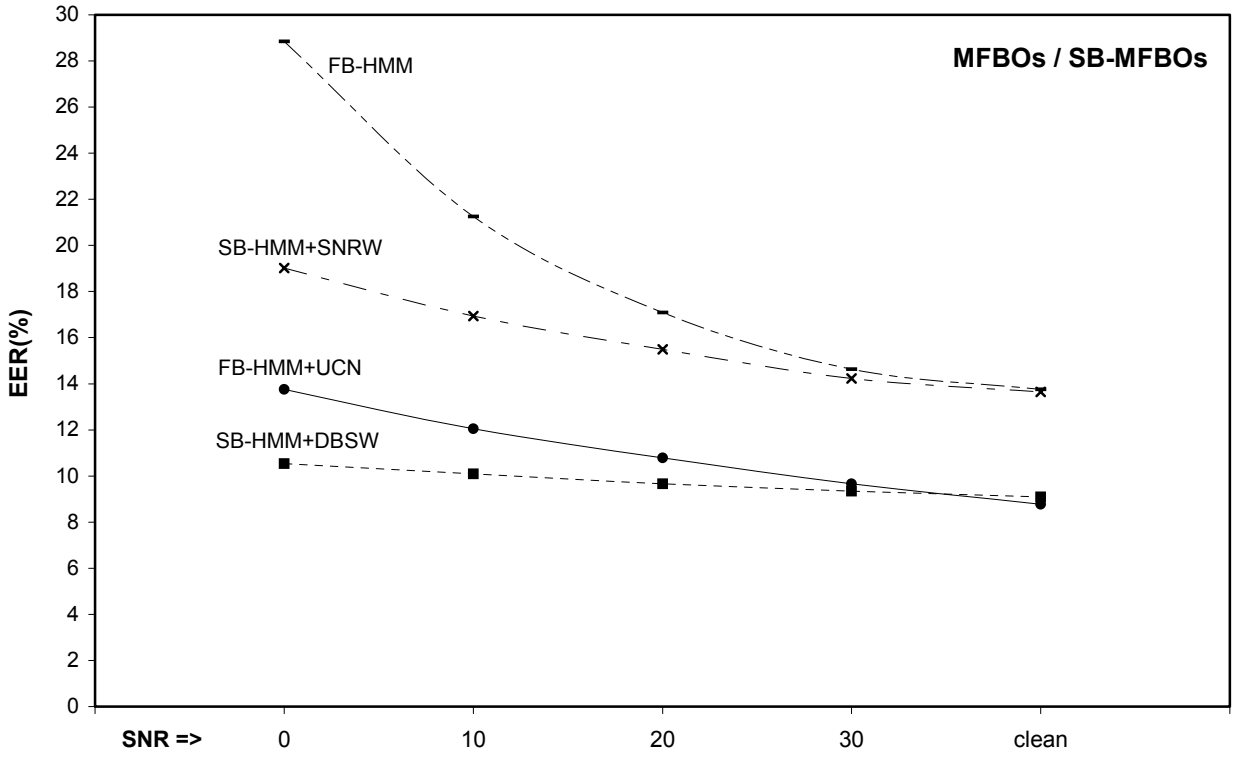


Figure 8: EER as a function of SNR for four different approaches. The feature parameters used in this study were MFBOs and SB-MFBOs.

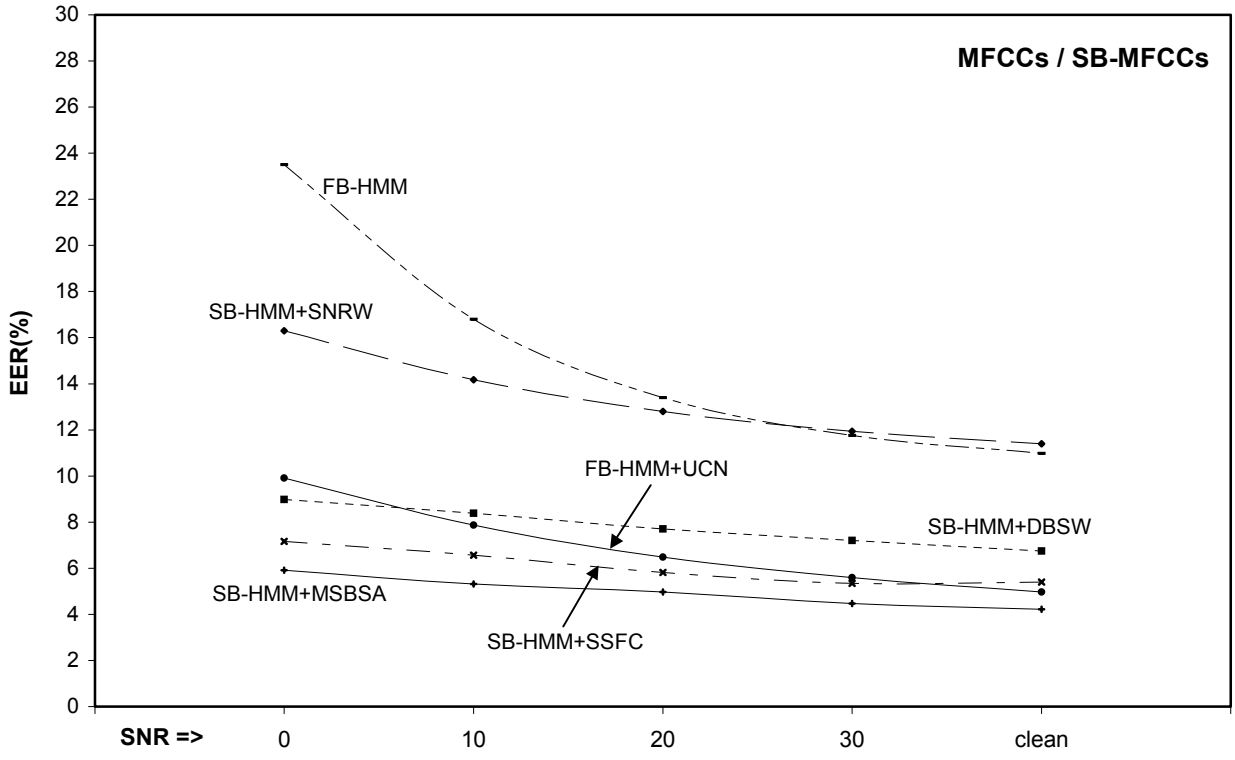


Figure 9: Relative performance of six different verification schemes as a function of SNR using MFCCs / SB-MFCCs.

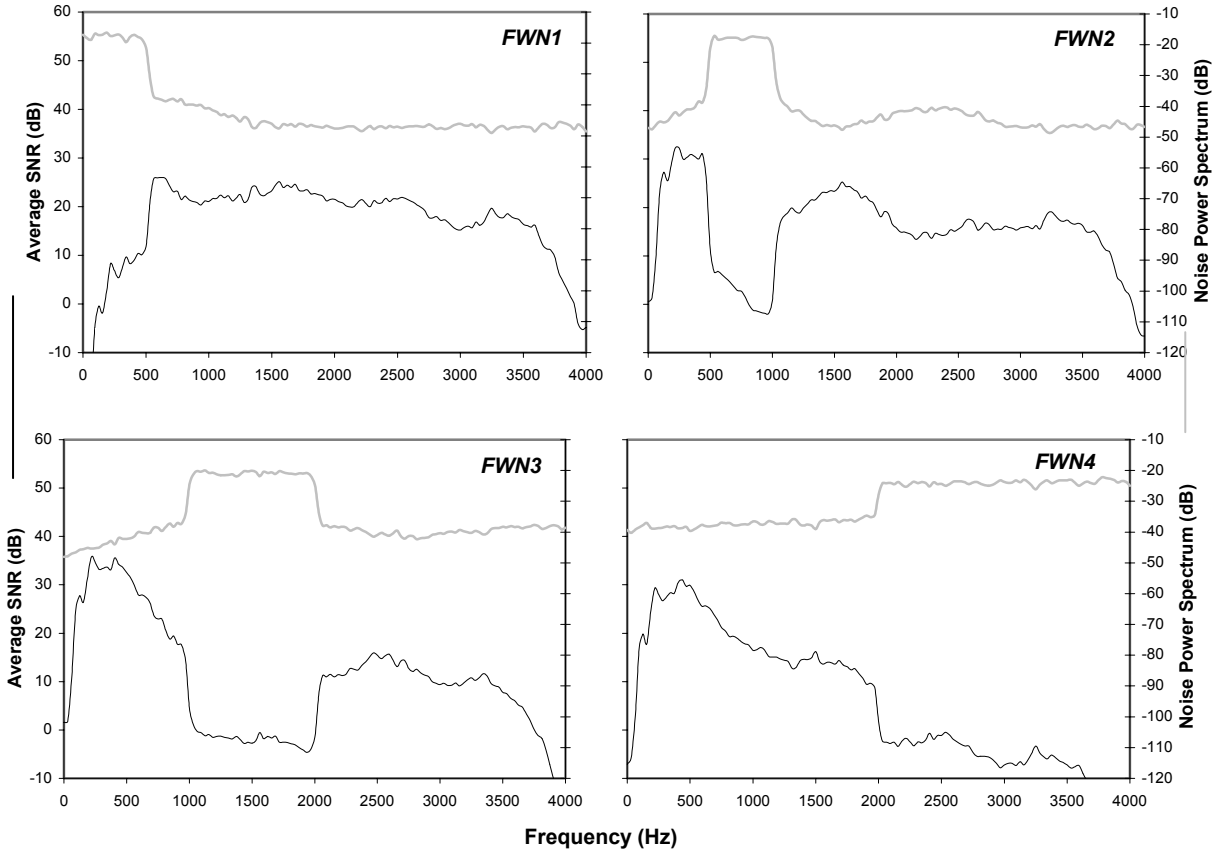


Figure 10: Power spectra of four types of narrow band noise and their affect on the average SNR (In *FWNn*, *FWN* stands for filtered white noise and *n* implies the association of the frequency region $\mathcal{A}_n$).

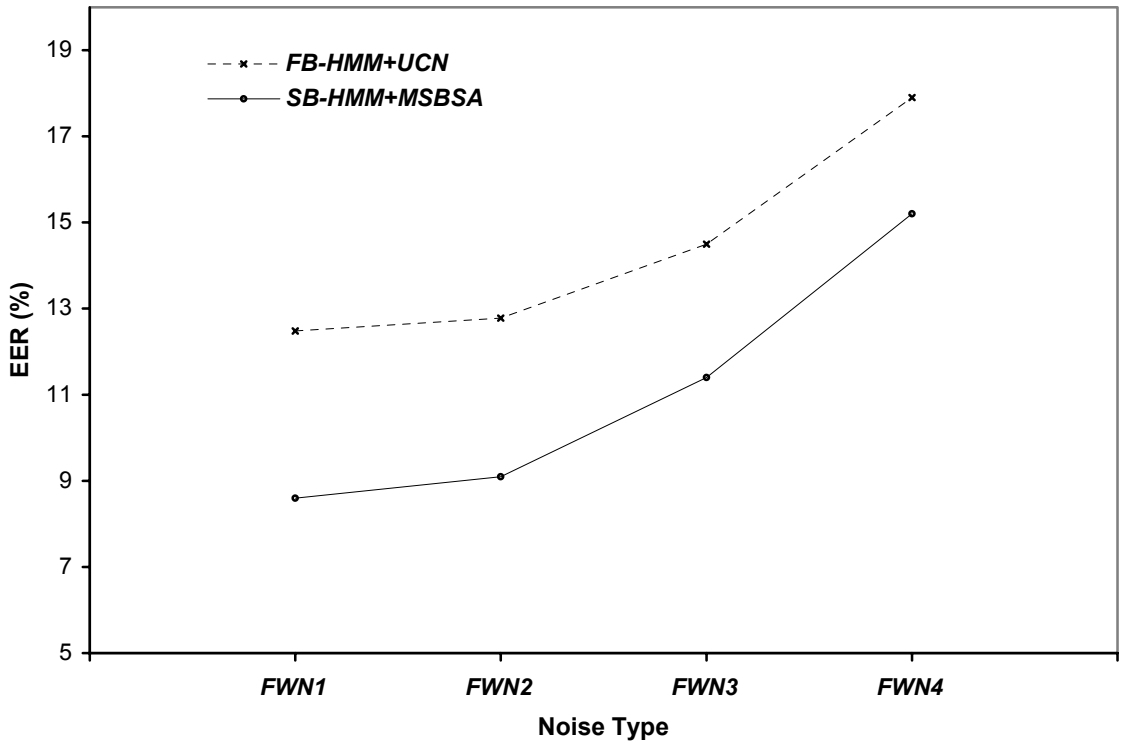


Figure 11: Relative effectiveness of the considered MSBSA-based method for speaker verification as a function of additive narrow band noise.

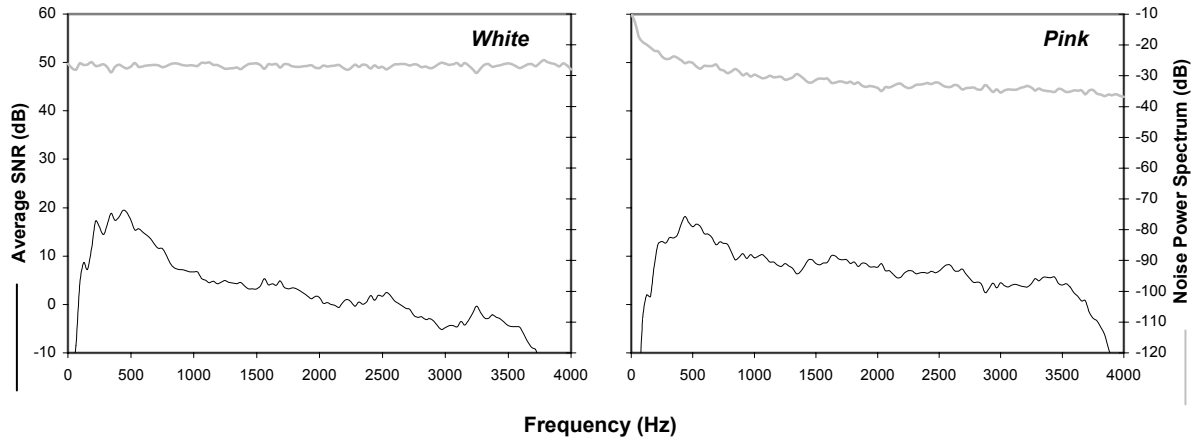


Figure 12: Power spectra of two types of noise and their effects on the average SNR.

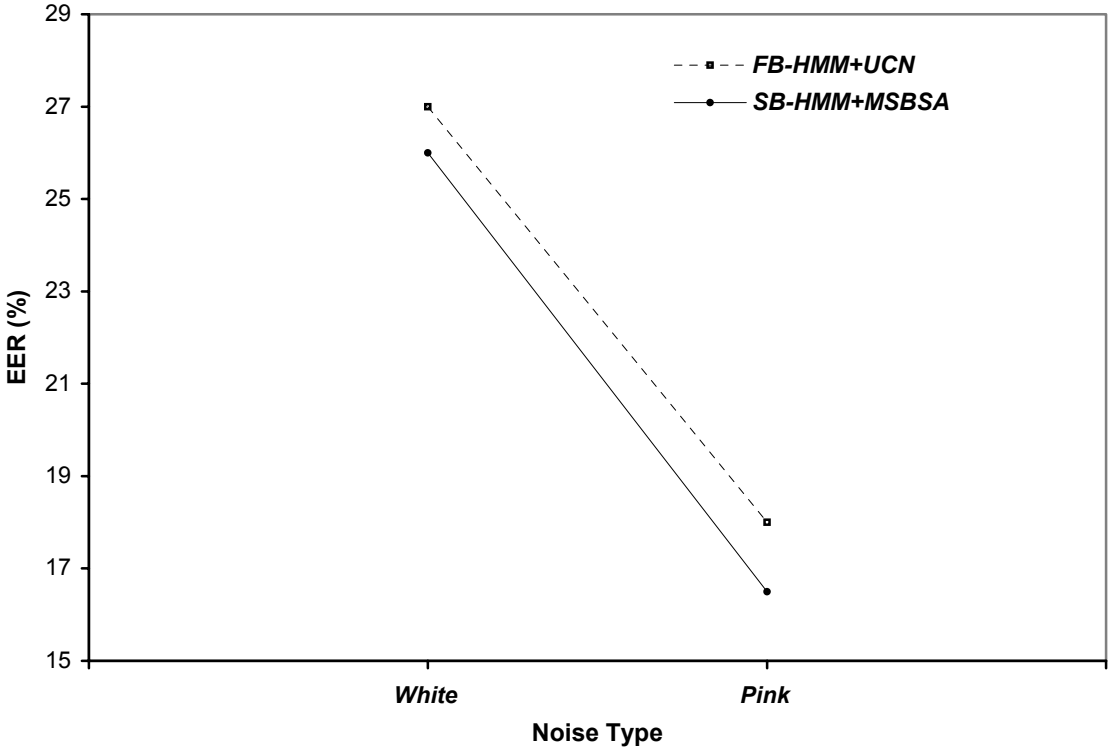
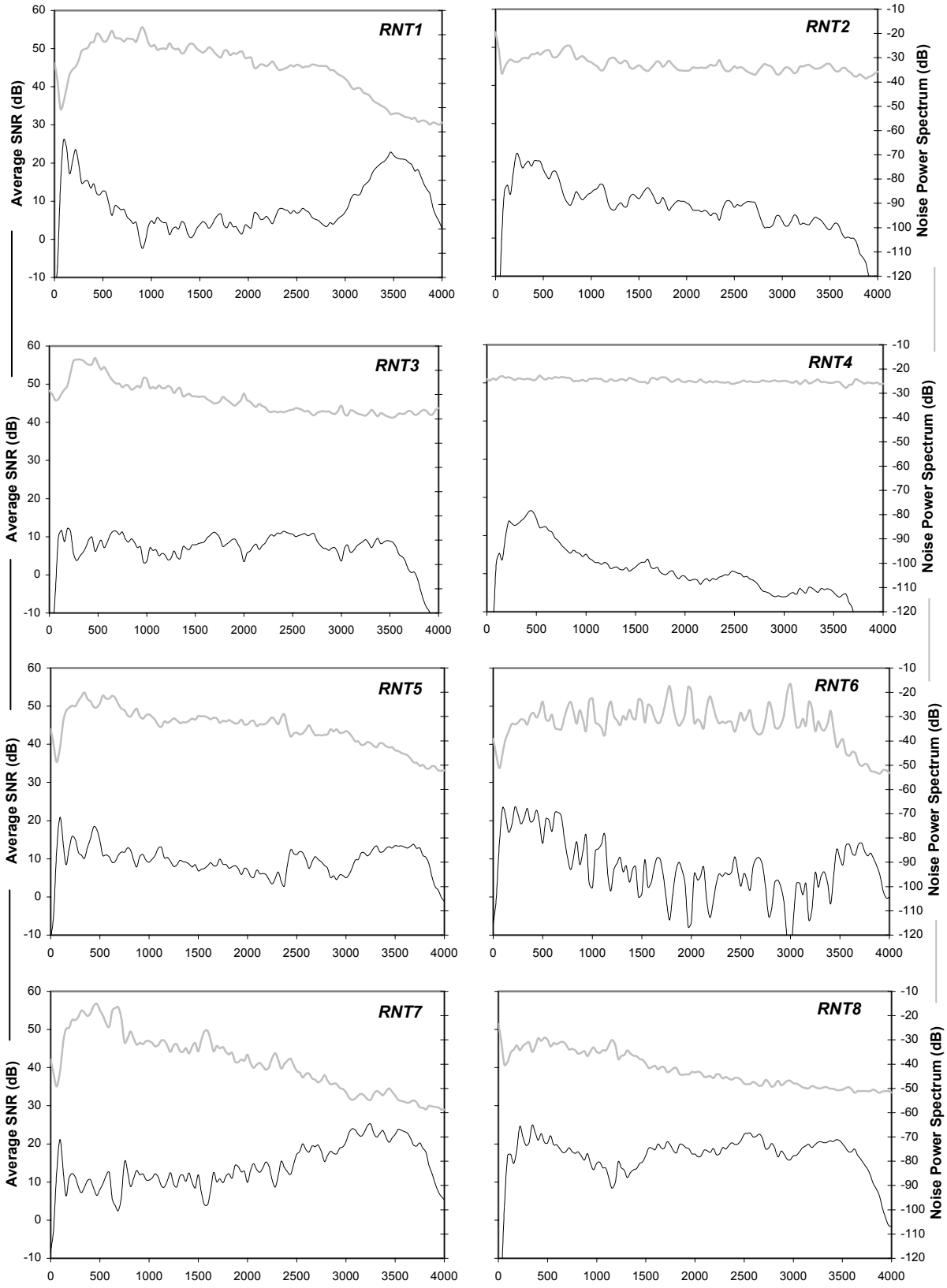


Figure 13: EERs for two different verification methods as a function of two types of additive noise.


RNT1
RNT2
RNT3
RNT4
RNT5
RNT6
RNT7
RNT8
Average SNR (dB)
Noise Power Spectrum (dB)
60
50
40
30
20
10
0
-10
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
0
500
1000
1500
2000
2500
3000
3500
4000

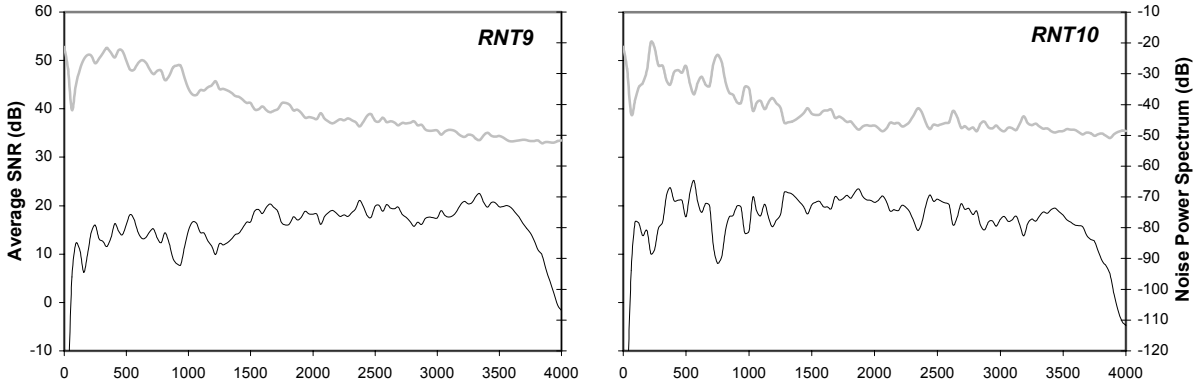


Figure 14: Power spectra of different types of real noise and their effects on the average SNR (In *RTNz*, *RTN* stands for real noise type and *z* is the associated noise type number as given on page 25).

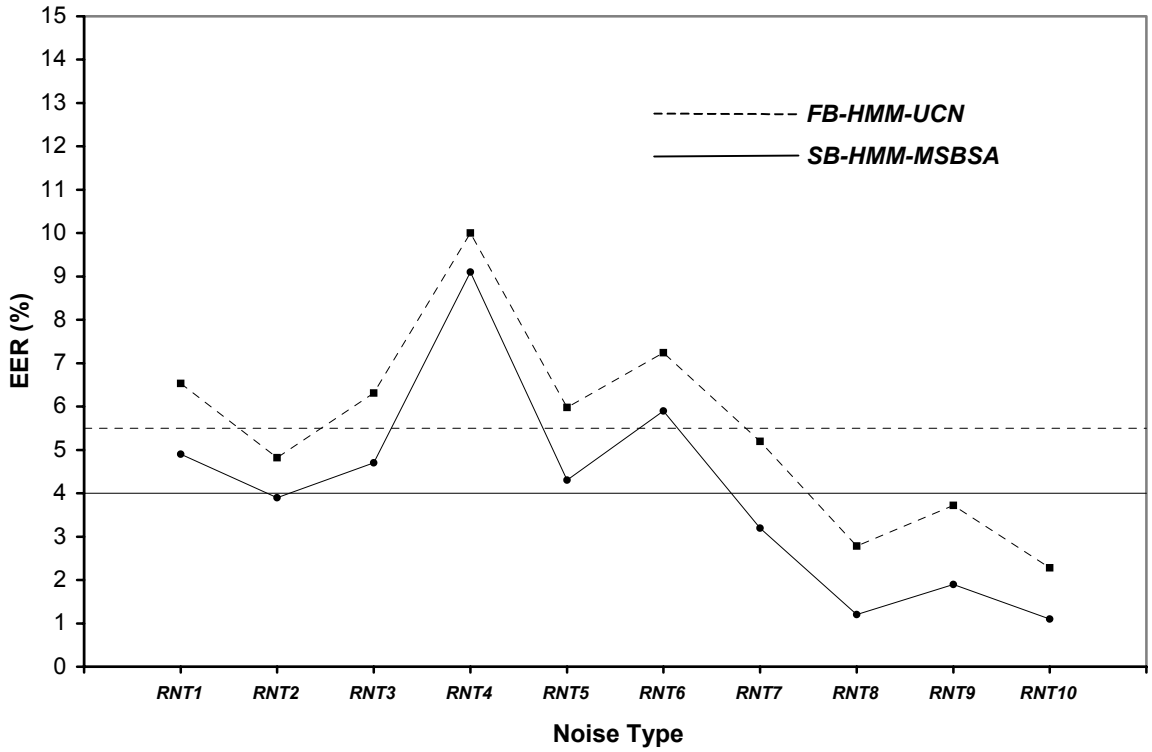


Figure 15: Results of the experiments carried out by contaminating the test utterances with different types of real noise (the flat lines represent the corresponding overall EERs).

Merging level	Single-frame	Phoneme	Word
EER (%)	13.8	14.7	15.4

Table 1: EERs in verification tests for different recombination levels.