Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 543–565

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition

Md. Sahidullah *, Goutam Saha

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, Kharagpur 721 302, India

Received 18 April 2011; received in revised form 14 November 2011; accepted 18 November 2011
Available online 26 November 2011

Abstract

Standard Mel frequency cepstrum coefficient (MFCC) computation technique utilizes discrete cosine transform (DCT) for decorrelating log energies of filter bank output. The use of DCT is reasonable here as the covariance matrix of Mel filter bank log energy (MFLE) can be compared with that of highly correlated Markov-I process. This full-band based MFCC computation technique where each of the filter bank output has contribution to all coefficients, has two main disadvantages. First, the covariance matrix of the log energies does not exactly follow Markov-I property. Second, full-band based MFCC feature gets severely degraded when speech signal is corrupted with narrow-band channel noise, though few filter bank outputs may remain unaffected. In this work, we have studied a class of linear transformation techniques based on block wise transformation of MFLE which effectively decorrelate the filter bank log energies and also capture speech information in an efficient manner. A thorough study has been carried out on the block based transformation approach by investigating a new partitioning technique that highlights associated advantages. This article also reports a novel feature extraction scheme which captures complementary information to wide band information; that otherwise remains undetected by standard MFCC and proposed block transform (BT) techniques. The proposed features are evaluated on NIST SRE databases using Gaussian mixture model-universal background model (GMM-UBM) based speaker recognition system. We have obtained significant performance improvement over baseline features for both matched and mismatched condition, also for standard and narrow-band noises. The proposed method achieves significant performance improvement in presence of narrow-band noise when clubbed with missing feature theory based score computation scheme.

Keywords: Speaker recognition; MFCC; DCT; Correlation matrix; Decorrelation technique; Linear transformation; Block transform; Narrow-band noise; Missing feature theory

1. Introduction

Speaker recognition is a biometric authentication process where the characteristics of human voice are used as the attribute (Kinnunen and Li, 2010; Campbell et al., 2009). A state-of-the art speaker recognition system has three fundamental sections: a feature extraction unit for representing speech signal in a compact manner, a modeling scheme to characterize those features using statistical approach (Campbell, 1997), and lastly a classification scheme for characterizing the unknown utterance. Most of the feature extraction techniques use low level spectral information which conveys vocal tract characteristics. The spectral information is extracted from 20–30 ms of speech signal using squared magnitude of discrete Fourier transform (DFT). As vocal tract is a slowly varying system, speech signal is nearly stationary over this analysis window. Hence, DFT based spectrum estimation technique is quite suitable. A systematic study of various spectral features can be found in (Kinnunen, 2004). Out of all existing features, *Mel frequency cepstral coefficient* (MFCC) is the

* Corresponding author. Tel.: +91 3222 283556/1470; fax: +91 3222 255303.
E-mail addresses: sahidullahmd@gmail.com (Md. Sahidullah), gsaha@ece.iitkgp.ernet.in (G. Saha).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2011.11.004

most popular and has become standard in speaker recognition system. MFCC is popular also due to the efficient computation schemes available for it and its robustness in presence of different noises.

In MFCC computation process, the speech signal is passed through several triangular filters which are spaced linearly in a perceptual Mel scale. The Mel filter bank log energy (MFLE) of each filters are calculated. Finally, cepstral coefficients are computed using linear transformation of MFLE. The linear transformation is essential here. The major reasons are as follows: (a) *improving the robustness*: the MFLEs are not much robust. They are very much susceptible to a small change in signal characteristics due to noise and other unwanted variabilities, (b) *decorrelation*: the log energy coefficients are highly correlated whereas uncorrelated features are preferred for statistical pattern recognition systems, specially for diagonal covariance based *Gaussian mixture model* (GMM) which is employed in today's speaker recognition system.

Amongst all linear transformation discrete cosine transform (DCT) is most popular and widely used for MFCC computation. The motivations behind the usage of DCT can be stated as follows. Firstly, the DCT is the sub-optimal approximation of the basis function of Karhunen–Loéve transform (KLT) when the correlation matrix of the sample closely approximates the correlation matrix of Markov-I process (Ahmed et al., 1974). The correlation matrix of MFLE data is fairly similar to the correlation matrix of first order Markov process. Secondly, DCT has the best energy compaction property for arbitrary data length compared to DFT and other sinusoidal transform like discrete sine transform (DST), discrete Hadamard transform (DHT), etc. (Oppenheim and Schafer, 1979). Though DCT based MFLE transformation technique is very popular, some studies have been carried out recently on further processing schemes of cepstral coefficient to improve the robustness against channel and other variabilities (Garreton et al., 2010; Hung and Wang, 2001; Nadersharif and Akbari, 2007). *Principal component analysis* (PCA) (Takiguchi and Ariki, 2007), *linear discriminant analysis* (LDA) (Kajarekar et al., 2001), *independent component analysis* (ICA) (Kwon and Lee, 2004), etc. are some traditional techniques which are also applied for formulating decorrelated features for speech processing applications.

Our proposed work is focused to design a linear transformation technique which can effectively preserve speech related information to improve the speaker recognition performance. Being motivated by the fact that the block wise filter bank outputs are more suitable for transformation using DCT, we have investigated block based transformation approach in case of traditional full-band based DCT which is applied to all the MFLE at a time. Earlier block based cosine transform has been applied for speech recognition (Jingdong et al., 2000). Recently, DCT is applied in a distributed manner (Sahidullah and Saha, 2009) to formulate feature for speaker identification. In image processing applications, DCT has also been applied in blocked manner (Jain, 2010). Subband DCT based coding method has been shown to be effective in image coding, image resizing schemes where DCT is computed for different block of subband (Jung et al., 1996; Mukherjee and Mitra, 2002). Here the signal is first divided into two parts: a high pass and a low pass and DCT is computed for each signals separately. On the other hand, subband based speaker recognition technology also gained attention as an alternative of conventional MFCC. In (Sivakumaran et al., 2003), different experimental results are reported based on subband DCT. During the last decade, several works have been carried out in subband processing based speaker recognition (Besacier and Bonastre, 2000; Finan et al., 2001; Damper and Higgins, 2003; Vale and Alcaim, 2008). The mathematical relationship between multi-band and full-band based MFCC coefficient are established in (Mak, 2002). In (Kim et al., 2008), subband DCT based MFCC is shown to perform better than full-band MFCC for different additive noises. There exists a number of other work on subband DCT or multi-band MFCC where it is shown to outperform existing baseline MFCC specially for partially corrupted speech signal (Besacier and Jean-Frantois, 1997; Ming et al., 2007; Jingdong et al., 2004). Though, it has played an effective role in improving performance of speech processing applications still multi block DCT is not much used in state-of-the art speaker recognition system. The main reason is that most of the existing works are at experimental level and the design issues related to multi-block configuration (i.e. number of bands, size of band, etc.) are yet to be precisely addressed. This is one of the main issue behind its unpopularity in spite of its superior empirical performance for speech and speaker recognition.

In our present work, the design issues related to block based MFCC computation scheme is addressed carefully along with a thorough experimental evaluation. The cepstral coefficient using multi-block DCT approach is systematically formulated. The scheme is also restructured for improving the performance of speaker recognition. The block transform (BT) based approach is shown to carry several levels of information. A novel block based approach is also proposed which has complementary information to the formerly proposed methods. The strengths of both the systems are combined using weighted linear fusion to get better performance. We have evaluated the performance of speaker recognition system with NIST SRE 2001 (for matched condition) and NIST SRE 2004 (for both matched and mismatched condition). The experimental result shows the superiority of our proposed block-based MFCC computation scheme for both the databases. As a final point, the paper proposes a technique where significant performance improvement is obtained for multi-block approach using linear transformation only. The system also performs better than standard MFCC for different types of noise. Additionally this system is significantly better than baseline system in case of narrow-band noise when

missing feature theory is applied (Lippmann and Carlson, 1997) by considering the scores of reliable and non-reliable feature with unequal degree.

The rest of the paper is organized as follows. First, in Section 2, a brief overview of MFCC computation is depicted mathematically for the completeness and better readability of the paper. In Section 3, different multi-block transformation techniques are formulated. In Section 4, several issues of block transformation techniques are discussed. Section 5 consists of the experimental setup, results and the discussion. Finally, the paper is concluded in Section 6 with some future directions.

2. MFCC computation: a matrix based approach

MFCC computation technique is based on DFT magnitude of speech frame. A detailed description of this process with block diagram can be found elsewhere (Chakroborty, 2008). In this work, we are analyzing the MFCC from mathematical point of view with the help of matrix operation notations. Though the different steps of MFCC calculation are standard and well-known, we review them briefly in a new way with a matrix based approach to formulate the problem addressed here.

Let, we have T number of speech frames each of size N extracted from a speech utterance. The followings are the different steps of MFCC computation.

1. *Windowing*: In the first stage, the signal is multiplied with a tapered window (usually Hamming or Hanning window). The windowed speech frames are given by,
$$[\boldsymbol{s}_w]_{T\times N} = [\boldsymbol{s}]_{T\times N} \circ [\boldsymbol{w}]_{T\times N}, \tag{1}$$
where $\boldsymbol{s}$ is a matrix containing framed speech, $\boldsymbol{w}$ is another matrix whose T rows contain same window function w of size N and $\circ$ denotes entry wise matrix multiplication.
2. *Zero-padding*: Zero-padding is required to compute the power spectrum using *fast Fourier transform* FFT. Sufficient numbers of zeros are padded using the following matrix operation:
$$[\boldsymbol{s}_{zp}]_{T\times M} = [\boldsymbol{s}_w]_{T\times N}[\boldsymbol{I} \quad \boldsymbol{O}]_{N\times M}, \tag{2}$$
where $\boldsymbol{I}$ is an identity matrix of size $N \times N$ and $\boldsymbol{O}$ is a null matrix of size $N \times (M - N)$. Here M is power of two and is greater than N.
3. *DFT computation*: The windowed speech frames are multiplied with twiddle factor matrix ($\boldsymbol{W}$) to formulate DFT coefficients ($\boldsymbol{\Omega}$). Half of the twiddle factor matrix is sufficient due to the conjugate symmetric property of Fourier transform. This operation can be expressed as,
$$[\boldsymbol{\Omega}]_{T\times\frac{M}{2}} = [\boldsymbol{s}_{zp}]_{T\times M}[\boldsymbol{W}]_{M\times\frac{M}{2}}. \tag{3}$$
4. *Power spectrum computation*: Power spectrum ($\boldsymbol{\Theta}$) is computed by entry wise multiplying the DFT coefficents with its conjugate. This can be written as,
$$[\boldsymbol{\Theta}]_{T\times\frac{M}{2}} = [\boldsymbol{\Omega}_w]_{T\times\frac{M}{2}} \circ \left[\boldsymbol{\Omega}_w^*\right]_{T\times\frac{M}{2}}. \tag{4}$$
5. *Filter bank log energy computation*: The speech signal is passed through a triangular filter bank of frequency response ($\boldsymbol{\Lambda}$) which contains p filters, linearly spaced in Mel scale. The log energy output ($\boldsymbol{\Psi}$) of the filter bank is given by,
$$[\boldsymbol{\Psi}]_{T\times p} = \log\left[[\boldsymbol{\Theta}]_{T\times\frac{M}{2}}[\boldsymbol{\Lambda}]_{\frac{M}{2}\times p}\right]. \tag{5}$$
6. *DCT computation*: In the finishing stage of MFCC computation, $\boldsymbol{\Psi}$ is multiplied with the DCT matrix $\boldsymbol{D}$ to create final co-efficient ($\boldsymbol{x}$). Therefore,
$$[\boldsymbol{x}]_{T\times p} = [\boldsymbol{\Psi}]_{T\times p}[\boldsymbol{D}]_{p\times p}, \tag{6}$$
where, each column of $\boldsymbol{D}$ are p-dimensional orthogonal basis vector of DCT. However, since the first coefficient is discarded as it is *dc-coefficient*, multiplication with a $p \times (p-1)$ matrix is adequate in DCT computation.

In this work, we have investigated a better alternative of standard DCT (i.e. $\boldsymbol{D}$) using block transform which is shown to give better speaker recognition performance with GMM-UBM based speaker recognition system.

3. Block transform approach in MFCC computation

Block based transformation are very popular in image coding (Akansu and Haddad, 1992). In this approach, the whole signal is divided into non-overlapping blocks and individual blocks are processed independently.

Let $\boldsymbol{F}$ be a signal matrix of dimension $T \times p$. Now in block transformation approach $\boldsymbol{F}$ is transformed with a linear kernel $\boldsymbol{L}$ of size $p \times d$ such that $\boldsymbol{L}$ is strictly a band matrix and it can be expressed as,

$$\mathrm{L} = \begin{bmatrix} \left.\begin{matrix}\cdot\\\cdot\\\cdot\\\cdot\end{matrix}\right\}\Phi_1 & \begin{matrix}\mathrm{O}\\\mathrm{O}\end{matrix} & \begin{matrix}\mathrm{O}\\\mathrm{O}\\\mathrm{O}\\\mathrm{O}\end{matrix} \\ \begin{matrix}\mathrm{O}\\\mathrm{O}\\\mathrm{O}\\\mathrm{O}\end{matrix} & \left.\begin{matrix}\cdot\\\cdot\\\cdot\\\cdot\end{matrix}\right\}\Phi_2 & \left.\begin{matrix}\cdot\\\cdot\\\cdot\\\cdot\end{matrix}\right\}\Phi_3 \end{bmatrix} \tag{7}$$

where $\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$ and $\boldsymbol{\Phi}_3$ are orthogonal matrices. This is the fundamental idea behind block transformation (BT) which is applied here for computing cepstral vectors from MFLE. In a standard BT, $\boldsymbol{\Phi}$-matrix is selected as an orthogonal transformation like DCT, DST, DHT, etc. The eigenvector of underlying covariance matrix of the signal vector is used to choose the orthogonal transformation. In MFCC computation, as the BT is applied on the MFLE data which are highly correlated and eventually follow Markov-I property, hence, DCT matrix is a better choice for $\boldsymbol{\Phi}$. The decomposition of Mel filter bank output into blocks is an important issue at this point. In image processing application, a large image is divided into smaller blocks of size 8×8 or 16×16. In speaker recognition system, the number of outputs from filter bank are not so large compared to the size

of images in image processing applications. In our proposed work, we are considering filter bank consists of 20 filters, which is mostly used in MFCC computation. Hence, we will be considering two or three block based approach. In experimental section (Section 5), we will observe that this choice is reasonably good for the given filter bank size. However, multi-block approach with more number of blocks could be used for the case where higher number of critical bands are considered. In the following subsections, we proposed three kinds of block transformation. The first form of BT which is based on crisp partitioning of MFLE, is referred as *Non-Overlapped Block Transform* (NOBT). On the other hand, *Overlapped Block Transform* (OBT) is the second category of BT which is formulated by extending the block sizes of NOBT in the direction of adjacent blocks. The last kind of BT introduced in this work is a special case of OBT where the basis functions of transformation are shifted form of each other. This transformation is named as *Shifted Basis Block Transform* (SBT) and it is shown to carry localized spectral information which is lacking in NOBT and standard OBT.

3.1. Non-Overlapped Block Transformation (NOBT)

The most elementary block transformation scheme is non-overlapped block transformation. In this scheme, the transformation matrix is direct sum of two orthogonal matrices, i.e. the transformation is carried out on two non-overlapping blocks. Therefore, the transformation matrix ($\boldsymbol{L}_{nobt}$) can be expressed as,

$$\boldsymbol{L}_{nobt} = \boldsymbol{\Phi}_1 \oplus \boldsymbol{\Phi}_2 \oplus \cdots \oplus \boldsymbol{\Phi}_N = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{O} & \dots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Phi}_2 & \dots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \dots & \boldsymbol{\Phi}_N \end{bmatrix}, \quad (8)$$

where $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \boldsymbol{\Phi}_3, \dots, \boldsymbol{\Phi}_N$ are orthogonal transformation matrices.

This design can be made such that the blocks will have equal size. In that case, the size of chunk must be a factor of total number of filters in the filter bank, i.e. p. In standard MFCC computation, DCT is used for its decorrelation property. We have experimentally observed that the nearer MFLEs (i.e. smaller blocks) are more suitable for DCT as they closely follow Markov-I property. This information is used for choosing $\boldsymbol{\Phi}_i$.

For example, let we consider two blocks of same sizes q such that $p = 2q$. Hence, the transformation matrix is given by,

$$\boldsymbol{L}_A = \begin{bmatrix} \boldsymbol{\Phi}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Phi}_q \end{bmatrix}, \quad (9)$$

where $\boldsymbol{\Phi}_q$ is a DCT matrix of size $q \times (q-1)$ and it is given by

$$\boldsymbol{\Phi}_q^{ij} = \sqrt{\frac{2}{q}} \cos\left[\frac{\pi i(2j+1)}{2q}\right]. \quad (10)$$

The dimension of transformation matrix for two block NOBT is $p \times 2(q-1)$ as total number of output co-efficient is $2(q-1)$ after discarding the dc-coefficients which have insignificant effect in speaker recognition. As $\boldsymbol{\Phi}_i$ is orthogonal hence, $\boldsymbol{L}_{nobt}^T \boldsymbol{L}_{nobt} = \boldsymbol{I}$.

The basis vectors and their frequency responses also have an interesting property for NOBT when each block has equal number of samples. For example, when the number of filter $p = 20$ then the basis vectors of DCT and the filter bank response of full-band DCT is shown in Figs. 1 and 2. The basis vector and frequency response of NOBT consisting of two blocks each of size *10* are shown in Figs. 3 and 4. Clearly the ith and $(i+10)$th basis functions (for $i = 1, 2, 3, \dots, 10$) of the NOBT are shifted basis pairs; hence, they have similar frequency response. The main lobe width of the NOBT is also higher than the full-band DCT based filters.

The cepstral coefficient ($\boldsymbol{x}^{nobt}$) using two block based NOBT can be written as,

$$\begin{aligned} \left\{\boldsymbol{x}_i^{nobt}\right\}_{i=1}^{q-1} &= \sqrt{\frac{2}{q}} \sum_{j=1}^{q} \boldsymbol{\Psi}(i) \cos\left[\frac{\pi i(2j+1)}{2q}\right], \\ \left\{\boldsymbol{x}_i^{nobt}\right\}_{i=q+1}^{p-q-1} &= \sqrt{\frac{2}{p-q}} \sum_{j=1}^{q} \boldsymbol{\Psi}(q+i) \cos\left[\frac{\pi (i-q)(2j+1)}{2(p-q)}\right]. \end{aligned} \quad (11)$$

The NOBT is a simple and computationally efficient scheme for calculating cepstral coefficient directly from MFLE. The choice of block size for NOBT is studied experimentally and discussed in Section 4.1. We have shown that formant specific block selection approach is better for speech feature computation. The advantages of NOBT over single transformation are: (i) NOBT has a localization effect. If speech spectrum is partially distorted due to several noises, then a part of the feature vector is only affected while the rest remain unaltered. (ii) As the dc-coefficient has less significant contribution in speaker recognition, the feature dimension of NOBT is lesser than that of full-band case where only one dc-coefficient can be discarded.

Despite the above listed advantages it has a major drawback due to its abrupt discontinuity in the boundary. In block based image coding schemes, this problem is addressed using lapped orthogonal transform (Malvar and Staelin, 1989). Motivated by this, in the subsequent subsection, a solution is prescribed using overlapping of neighborhood blocks.

3.2. Overlapped Block Transformation (OBT)

In this scheme of block transform, the neighborhood blocks share some filter bank log energy coefficients to avoid the discontinuity at the end. In Fig. 5, overlapped block transformation matrix ($\boldsymbol{L}_{obt}$) is shown with two blocks of block size q_a and q_b where the total number of elements in MFLE is p. $\boldsymbol{\Phi}_A$ and $\boldsymbol{\Phi}_B$ are two orthogonal

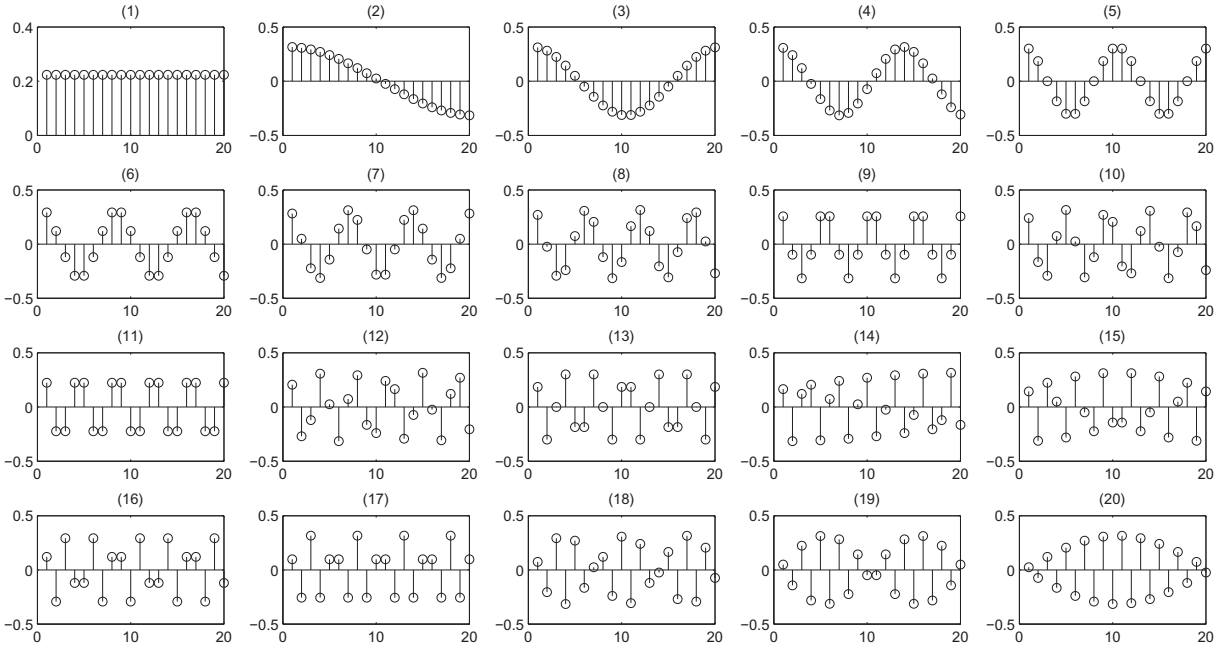

Fig. 1. Basis functions of DCT filter bank. The titles of the subplot indicate the sequence numbers of the basis functions.

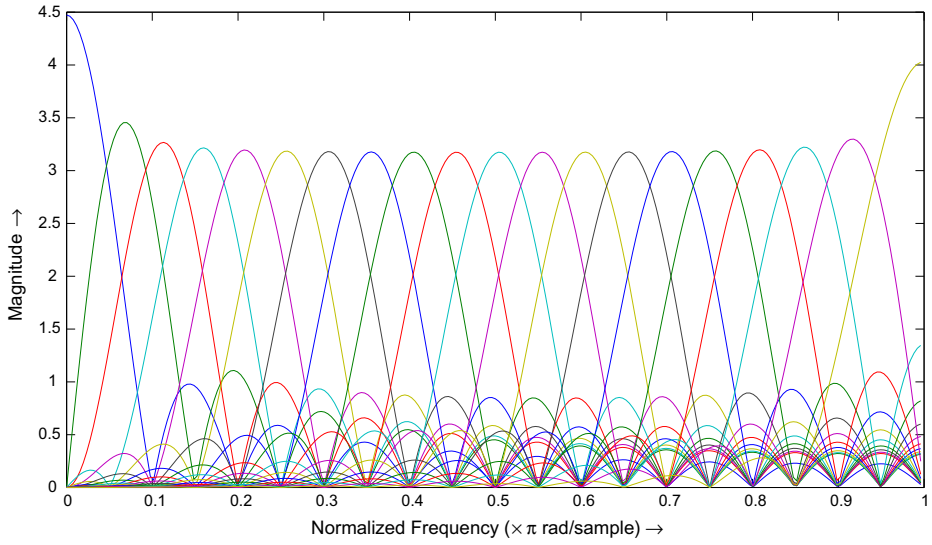

Fig. 2. Superimposed frequency responses of twenty filters of DCT filter bank of Fig. 1.

transforms, i.e. DCT matrices; and $\boldsymbol{O}_I$ and $\boldsymbol{O}_{II}$ are matrices of null elements.

The coefficients for OBT for two block of size q_a and q_b are defined as,

$$\left\{x_i^{obt}\right\}_{i=1}^{q_a-1} = \sqrt{\frac{2}{q_a}}\sum_{j=0}^{q_a-1}\boldsymbol{\Psi}(j)\cos\left[\frac{\pi i(2j+1)}{2q_a}\right],$$
$$\left\{x_i^{obt}\right\}_{i=q_a}^{q_a+q_b-1} = \sqrt{\frac{2}{q_b}}\sum_{j=0}^{q_b-1}\boldsymbol{\Psi}(p-q_b+j)\cos\left[\frac{\pi(i-q_a+1)(2j+1)}{2q_b}\right], \tag{12}$$

where the amount of overlap for the two blocks are given by $q_a + q_b - p$. If equal extension of each block is considered then each block is extended by $(q_a + q_b - p)/2$ samples.

The above scheme can also be extended for arbitrary number of blocks, of different size and overlap. If we assume that the number of elements in each block are equal and is an even number say, $2r$ and overlapping with the adjacent blocks are 50%, i.e. r, then the total number of blocks can be expressed as $\left(\frac{p}{r}-1\right)$. Hence, the total number of coefficients, $m = \left[\left(\frac{p}{r}-1\right)(2r-1)\right]$.

The OBT effectively captures local (in frequency domain) spectral information with localized transformation of MFLE. On the other hand, it generates less distorted coefficients for partially corrupted speech signal. However, OBT with multiple blocks and larger overlap has some

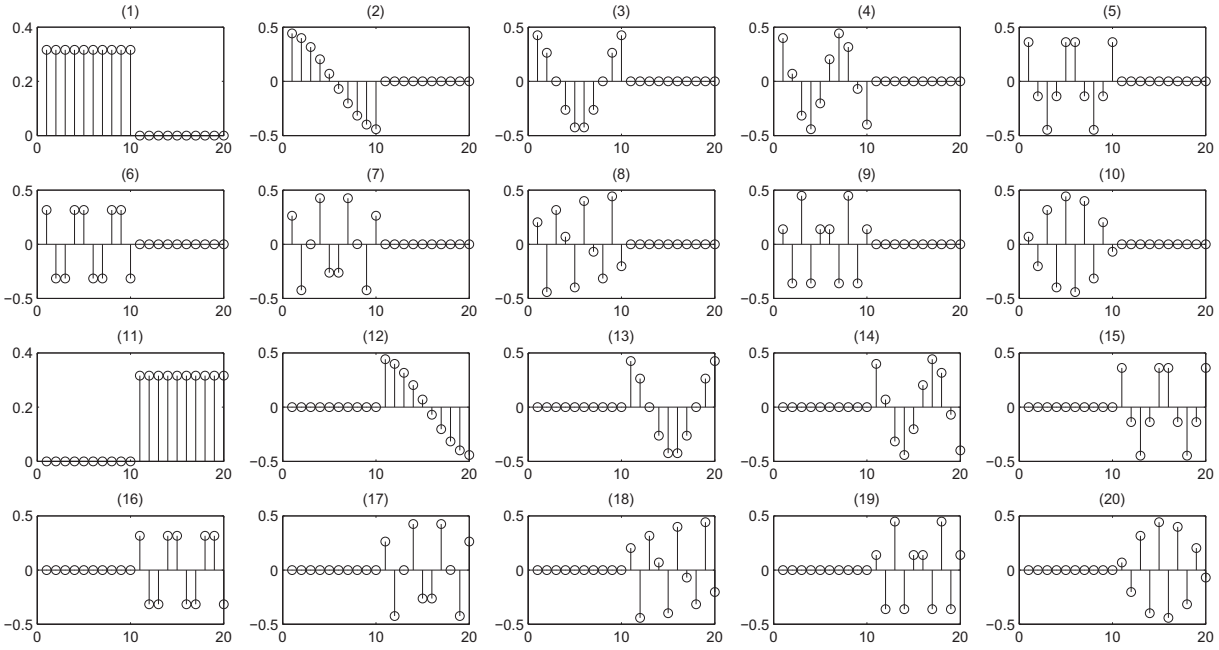

Fig. 3. Basis functions of block DCT with two non-overlapping blocks of equal size. The titles of the subplots indicate the sequence numbers of the basis functions.

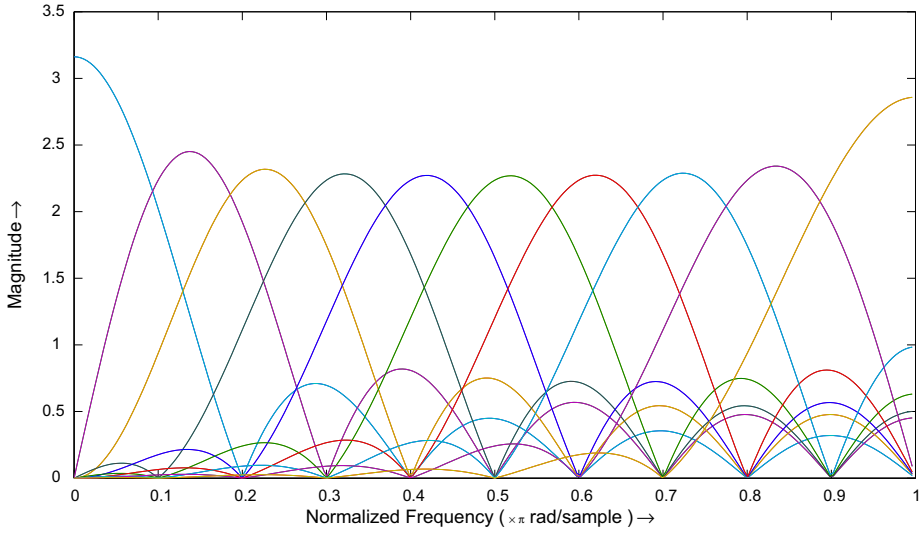

Fig. 4. Superimposed frequency responses of twenty filters of DCT filter bank of Fig. 3.

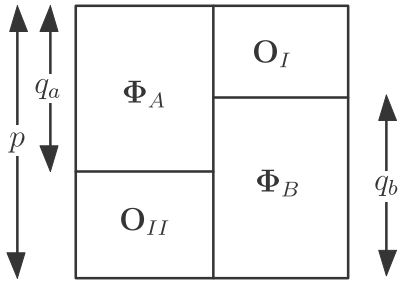

Fig. 5. Transformation matrix for OBT with two blocks of size q_a and q_b where the total number of MFLE is p. $\boldsymbol{\Phi}_A$ and $\boldsymbol{\Phi}_B$ are orthogonal transformations for two blocks. $\boldsymbol{O}_I$ and $\boldsymbol{O}_{II}$ are two null matrices.

disadvantages. Firstly, the dimension of feature increases significantly with the number of blocks. Secondly, its efficiency degrades for speech signal corrupted with noise that affects more number of blocks. In the next subsection, we propose a novel form of OBT which can be effectively used to improve the performance of existing system.

3.3. Shifted Basis Block Transformation (SBT)

In this section, we propose a new transformation which has the advantage of multi-block transform as well overlapped transform. This technique keeps transitional information of filter bank log energies as local detail of

MFLE. It is basically the difference between the log energies of the filters in filter bank. Earlier, spectral difference in frequency axis has been proposed in (Nitta et al., 2000) for speech recognition. It has been observed that the relative subband energies contain significant information for speech and speaker recognition (Chetouani et al., 2009). We have defined a shifted basis function which will capture this information in an effective manner.

Spectral difference between subbands is nothing but relative energy of the subbands which can be calculated by differentiating log energies. The transformation output for the proposed method can be written as,

$$\left\{x_i^{sbt}\right\}_{i=1}^{p-2} = \Psi(i) - \Psi(i+2). \tag{13}$$

As we are skipping one subband for computing the coefficients the total number of output will be $(p-2)$, and the transformation kernel ($\boldsymbol{L}_{sbt}$) can be expressed for $p = 10$ as follows:

$$\boldsymbol{L}_{sbt} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{14}$$

The weighting is proposed in such a manner that the basis function of the transformation matrix are orthogonal to each other. This transformation can be viewed as a filter bank where all the filter will have equal magnitude response (as shown in Fig. 6) with a single large main lobe. It is unlike DCT where it has narrow side lobes.

From Eq. (14) it is also very clear that this transformation is a special kind of multi-block overlapped transformation where each block is of size three and there is an overlap of two samples between the consecutive blocks. The SBT computation scheme is very similar to delta feature computation scheme in spatial domain. Hence, it contains transitional information of different frequency bands. The BTs proposed in the previous subsections deals with large segmental information. But the SBT contains more detail attributes which some how ignored by full-band and other block based transformations. Hence, SBT and block based information contains some amount of complementary information. Therefore, the advantages of both the feature can be used in combined system (Chakroborty and Saha, 2010; Sahidullah et al., 2010) where both the performances are fused together to get better speaker recognition result.

Table 1
Database description for speaker recognition experiments. The database details (i.e. target model, test segment, and trial information) are shown for core-test section.

Specification	NIST SRE 2001 (Przybocki and Martin, 2002)	NIST SRE 2004 (Martin and Przybocki, 2006)
No. of speakers	174	310
Speech format	8 kHz, μ-law	8 kHz, μ-law
Speech quality	Cellular phone	Various telephonic
Channel variability	No	Yes
Handset variability	Yes	Yes
Language variability	Yes	Yes
Number of target models	174 (Male: 74, female: 100)	616 (Male: 246, female: 370)
Number of test segments	2038	1174
Total trials	22418	26224
Correct trial	2038	2386
Impostor trial	20380	23838

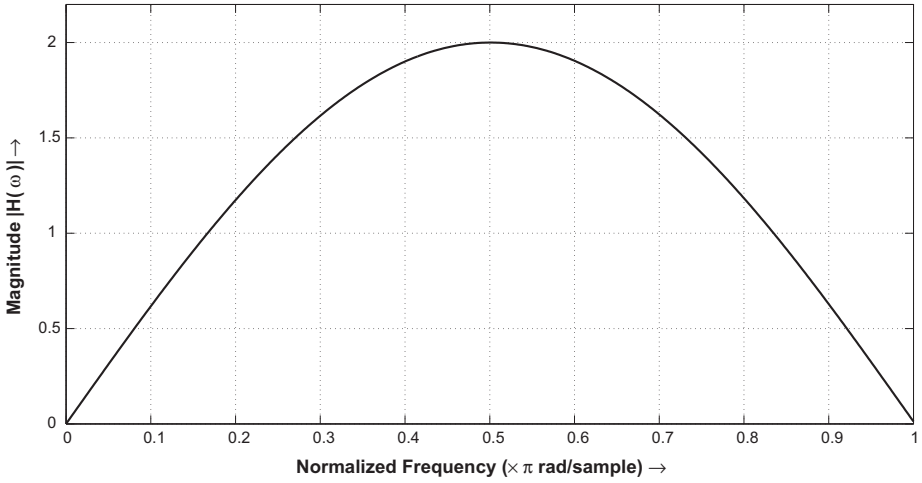

Fig. 6. Plot showing the frequency response of the SBT filter-bank. Each filter of the filter bank has equal frequency response ($|H(\omega)| = 2\sin\omega$).

4. Analysis of BTs

4.1. Decomposition of frequency bands for BT

In proposed block based transformation the filter bank log energies are decomposed into several blocks unlike standard full-band based technique. The decomposition is the most non-trivial part of this block based design. We need to decompose in such a way that most of the relevant information should be preserved and also it should reduce most of the superfluous effect. We make out that spectral peaks which includes formant frequencies contain most relevant speech and speaker related information. Each and individual formant frequencies also shows special attributes for different kinds of speech segments. In order to get an idea about the spectral peak distribution, we have plotted first three peaks for all the frames of a database. In Fig. 7(a), the distribution of spectral peaks for all the speech frames of NIST SRE 2001 (for male only) is shown. It is quite clear from the figure that spectral peaks are highly concentrated in certain frequency zone. Our block decomposition approach relies on this basic assumption that independent processing of these zones would effectively improve the speech parametrization process. By following this method, we can make sure that one spectral peak will not much affect other peaks while computing cepstral feature. The subband information is shown in Table 2 for 20 band case. It is also clear from the Fig. 7 that the first peak is mostly concentrated in frequency zone covered by first 8 filters. The rest of the area is jointly occupied by second and third peak. It gives moderately better decomposition for two band case if the size of the two blocks are chosen as 8 and 12. The frequency bands covered by the two bands are 0-883.17Hz and 745.93–4000 Hz which are dominated consecutively by first two formants F_1 and F_2, and they supposed to carry more speech related information (Douglas, 2009). Feature extracted through this way is called as NOBT-8-12 in this paper. On the other hand, we call equal size block based feature as NOBT-10-10 where each block is of size 10. For both of these cases, the feature dimension is 36 as we have further concatenated delta coefficients. We have extended NOBT-8-12 to overlapped form, and this feature is named as OBT-9-13. This OBT feature also effectively captures the frequency zone for first peak and combined zone of next two peaks. In the presence of noise, the high frequency critical bands are severely distorted. As a result, the spectrum is not correctly estimated. Hence, we keep only one block for F_2 and F_3 zone for OBT-9-13 feature.

Table 2
Subband frequency specification (in Hz) for Mel filter bank with 20 filters. f_{low} and f_{high} indicates the lower and upper frequency bound of the filters of triangular filter bank.

Subband no.	f_{low}	f_{high}	Subband no.	f_{low}	f_{high}
1	0	139.19	11	1033.43	1378.11
2	66.44	218.84	12	1197.97	1575.36
3	139.19	306.06	13	1378.11	1791.33
4	218.84	401.55	14	1575.36	2027.80
5	306.06	506.10	15	1791.33	2286.71
6	401.55	620.58	16	2027.80	2570.20
7	506.10	745.92	17	2286.71	2880.59
8	620.58	883.17	18	2570.20	3220.45
9	745.92	1033.43	19	2880.59	3592.57
10	883.17	1197.97	20	3220.45	4000

4.2. Decorrelation property of BT

One of our main motivations behind the use of BT is its decorrelation property. We have evaluated the *uncorrelatedness* of the transformed coefficients for different transformation kernel using the residual correlation measure (ϵ). It is defined as the mean of absolute value of all the off-diagonal elements of the correlation matrix and we define it for a transformation matrix $\boldsymbol{H}$ as,

$$\epsilon(\boldsymbol{H}) = \frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}|\boldsymbol{H}^T\boldsymbol{R}\boldsymbol{H} - \boldsymbol{I}|, \tag{15}$$

where $\boldsymbol{H}$ is of dimension $d' \times d$, $\boldsymbol{R}$ is the $d' \times d'$ correlation matrix of the data before transformation and $\boldsymbol{I}$ is $d \times d$ identity matrix. A transformation kernel, $\boldsymbol{H}$ will be suitable for decorrelation if it has lower residual correlation. In Fig. 8, the logarithm of residual correlation is shown for different transformation for different values of ρ for Markov-I process. It is desirable to note that though DCT is optimal for ideal Markov-I process, still the transformed coefficient for proposed BTs are substantially better than raw untransformed MFLE.

As noted earlier the correlation matrix of MFLE is not exactly same as Markov-I process. Hence, it is required to evaluate the decorrelation performance of the proposed BTs for real speech data. In Fig. 9, we have shown the residual correlation for different transformation on YOHO and POLYCOST databases for 50 randomly chosen speakers. In most of the cases, the residual correlation of the DCT is higher whilst for the BT based schemes the residual correlation is lesser. It is desired to note that for POLYCOST databases the speech signal is collected over telephone channel and the performance of BTs are comparatively better in this case. Recently, for the evaluation of state-of-the art speaker recognition systems, researchers use NIST SREs where the speech signal is collected over various kinds of channels. Hence, our technique could be effective for those data.

We then performed similar kind of experiment on the training portion and UBM of both the databases. The result is shown in Table 3. The proposed NOBT-10-10 and NOBT-8-12 are shown to decorrelate the MFLE in efficient manner than the standard DCT. The decorrelation property of OBT-9-13 is also better than that of DCT for NIST SRE 2004. However, the decorrelating property of SBT is not much efficient even like DCT, still it is better than untransformed MFLE data.

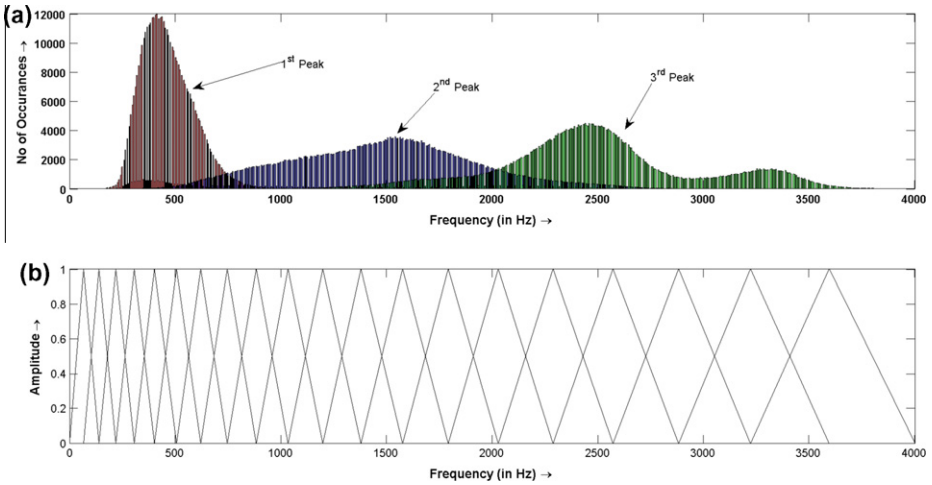

Fig. 7. Figure showing (a) superimposed histogram of first three spectral peaks, (Data are taken from the male section of NIST SRE 2001.) and (b) the Mel filter bank structure for 20 filters.

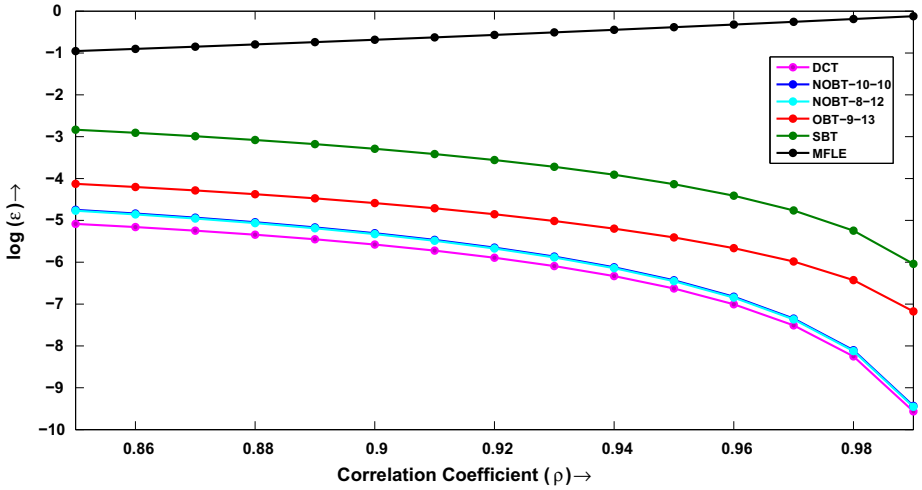

Fig. 8. Figure showing plot of logarithm of residual correlation of different transformations. The residual correlations are computed for different values of ρ where the correlation matrix of the given data follows ideal Markov-I property.

Table 3
Residual correlation (ϵ) for training data combined with the corresponding background data for different transformation. The last row shows the residual correlation of raw MFLE for both the databases.

Transform name	NIST SRE 2001	NIST SRE 2004
DCT	0.0725	0.0769
NOBT-10-10	0.0605	0.0615
NOBT-8-12	0.0574	0.0570
OBT-9-13	0.0744	0.0733
SBT	0.1452	0.1334
No transform	0.4397	0.4404

4.3. Effect of noise in BT based MFLE transformation

Speech signal is distorted both in time and frequency domain when contaminated by various kinds of noise. In spite of its overall degradation in presence of noise, several frequency bands remain unaltered or less affected. Using block based feature computation scheme, we may possibly obtain less distorted features as all the subband log energies are not used at a time for cepstral computation. In Figs. 10 and 11, we have shown the effect of additive white Gaussian noise and additive HF channel noise on the various proposed block based transforms. The noise samples are taken from NOISEX-92[1] database. We have down sampled the original noise with the sampling frequency 8 kHz and added to the speech signals. The effect of noise is shown for signal-to-noise ratio (SNR) of 15 dB on randomly selected speech frame. For very short duration of time (here 20 ms, i.e. 160 samples per frame) the noise affects different subbands with unequal strength. As a consequent several subbands remain less effected and others are highly distorted. In both the figures, this phenomena can be observed in Figs. 10(c) and 11(c) for MFLE. We can check

[1] http://www.spib.rice.edu/spib/select_noise.html.

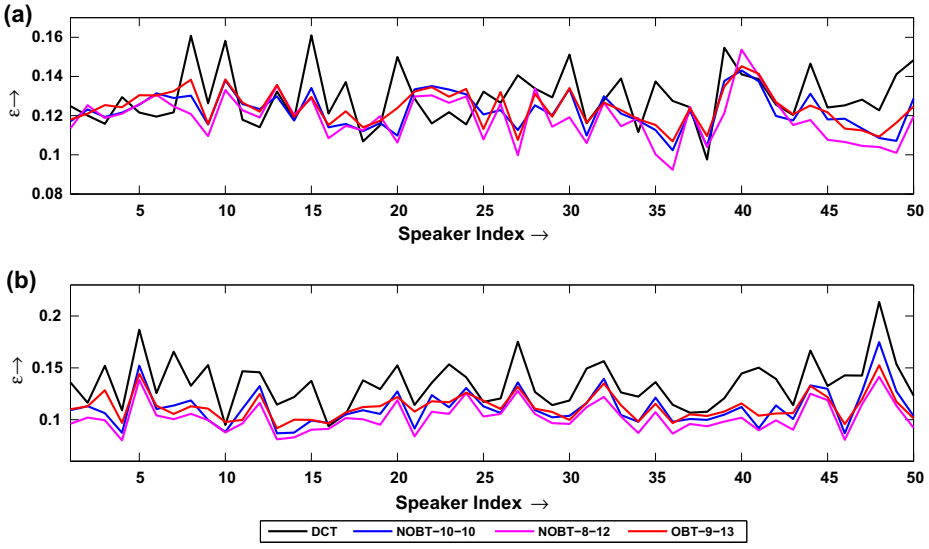

Fig. 9. Figure showing residual correlation of different transformations for practical speech data. The plots are shown for 50 randomly chosen speakers of (a) YOHO (microphonic) and (b) POLYCOST (telephonic) databases.

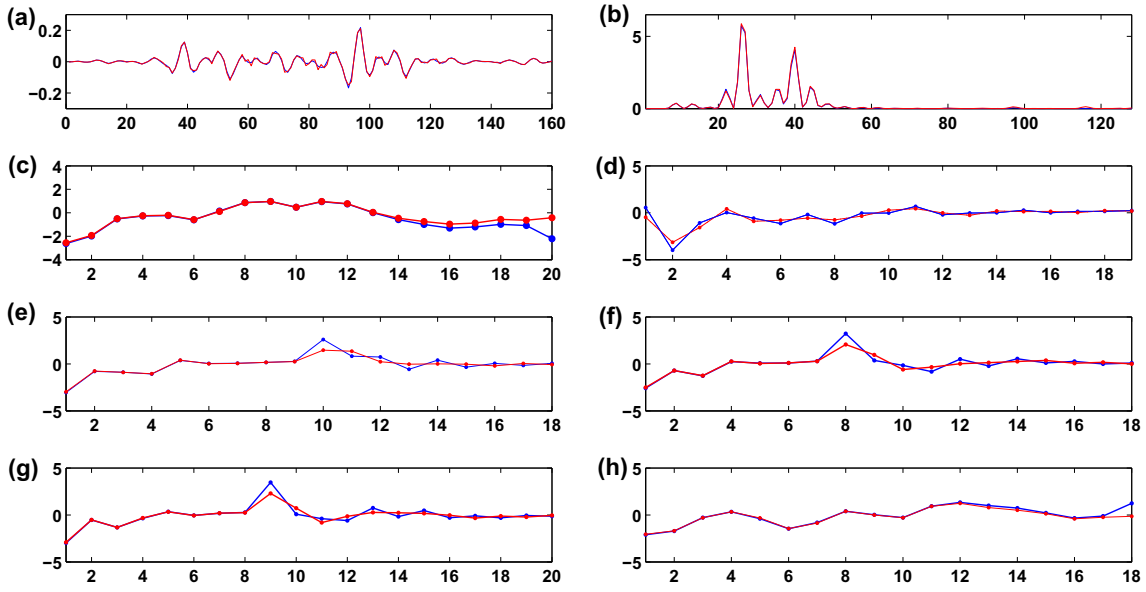

Fig. 10. Effect of additive white Gaussian noise (SNR:15 dB) on (a) speech signal, (b) power spectrum, (c) mel filter bank log energy. The effect is also shown for different cepstrum based on (d) DCT, (e) NOBT-10-10, (f) NOBT-8-12, (g) OBT-9-13, and (h) SBT. (Blue line: clean speech, red line: noisy speech.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

that a few MFLE remain almost unchanged even in presence of 15 dB SNR of noise while others are significantly changed. Therefore, if we apply full-band transformation like DCT more number of coefficients will be affected in presence of both types of noise. The subfigures of Figs. 10 and 11 interprets this by showing transformed coefficients based on DCT, NOBT-10-10, NOBT-8-12 and OBT-9-13.

The proposed BTs are much more efficient than standard DCT based approach when the speech signal is corrupted by narrow-band noise. We have observed the effect of narrow-band noise on speech signals and its transformed coefficients. In our work, this type of noise is synthetically generated using two methods. First method is to pass the white noise sample of NOISEX-92 database through a band pass filter with sharp bandwidth. The filter is designed using 6th order Butterworth approximation with lower and upper cut-off frequency as 2000 Hz and 2300 Hz, respectively. The second method is to add four frequency components (i.e. sinusoidal tones) of 2000 Hz, 2100 Hz, 2200 Hz and 2300 Hz. The amplitudes of the sinusoids are chosen randomly. The two narrow-band signals generated here is called as Type-I and Type-II narrow-band noise respectively. As Type-II narrow-band signal is generated by just adding sinusoidal we can call it more *pure*

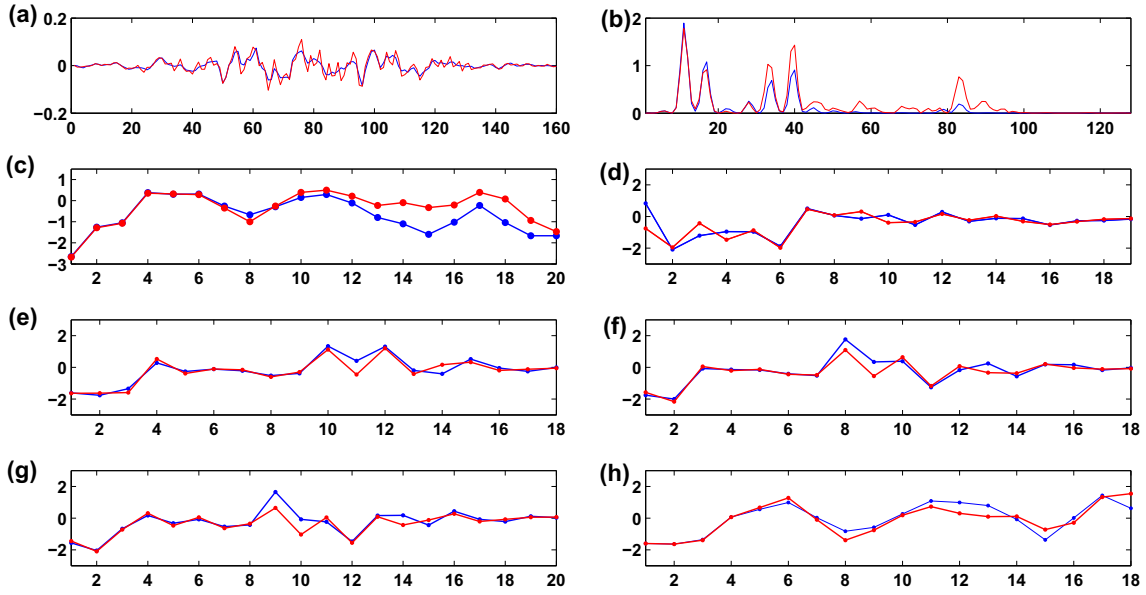

Fig. 11. Effect of hfchannel noise (SNR:15 dB) on (a) speech signal, (b) power spectrum, (c) mel filter bank log energy. The effect is also shown for different cepstrum based on (d) DCT, (e) NOBT-10-10, (f) NOBT-8-12, (g) OBT-9-13, and (h) SBT. (Blue line: clean speech, red line: noisy speech.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

narrow-band than Type-I. Hence, the effect of Type-I is more local than that of Type-II. The speech spectrogram affected by narrow-band noise is shown in Fig. 12. On the other hand, how the presence of narrow-band signal affects feature extraction scheme is shown in Fig. 13. It is noteworthy to mention that the cepstral features extracted from the affected zone gets severely affected. Conversely, features extracted from the other zone are almost unaltered. Therefore, we can get improved speaker recognition performance if we successfully select the unaffected features and ignore the distorted ones. This could be accomplished using missing feature theory (MFT) (Lippmann and Carlson, 1997). The experimental results using MFT based scoring scheme is presented in Section 5.2.6.

4.4. Computation complexity

The proposed BTs have another major advantage over exiting full-band based transformation due its low computational cost. In Fig. 14, the structure of various

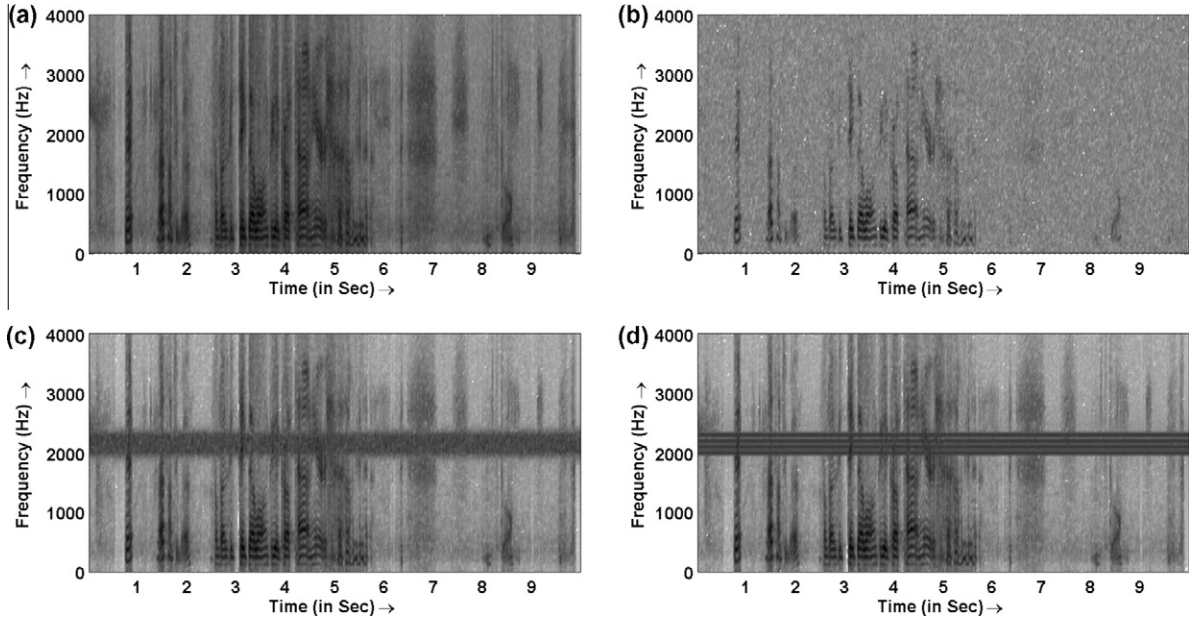

Fig. 12. Spectrogram showing characteristics of (a) clean speech signal, (b) effect of white noise, (c) effect of Type-I narrow-band noise, (d) effect of Type-II narrow-band noise. In all the cases SNR is set at 15 dB.

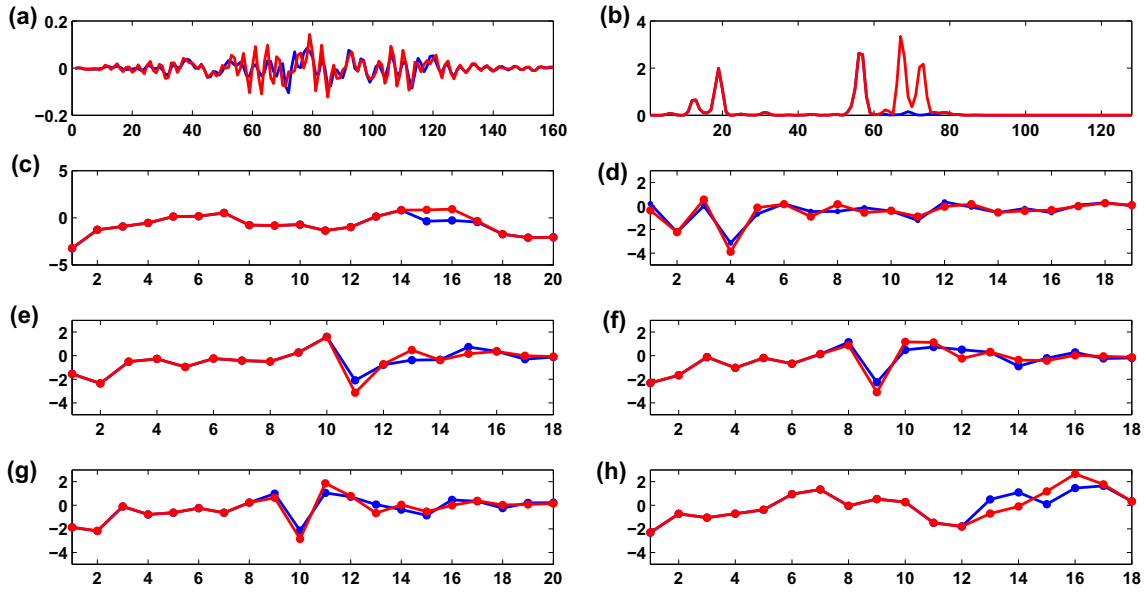

Fig. 13. Effect of narrow-band noise (Type-I, SNR:15 dB) on (a) speech signal, (b) power spectrum, (c) mel filter bank log energy. The effect is also shown for different cepstrum based on (d) DCT, (e) NOBT-10-10, (f) NOBT-8-12, (g) OBT-9-13, and (h) SBT. (Blue Line: Clean Speech, Red Line: Noisy Speech.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

transformation matrix is shown along with standard DCT based transformation. The BT matrices are sparse in nature, hence, the computational time is significantly less than full matrix based transformation.

For example, if we use filter bank of size 20, then total number of multiplication required for DCT is 380. While for NOBT-10-10 it is 180 and for NOBT-8-12 it is 188. On the other hand for OBT-9-13 and OBT-8-8-8, the required number of multiplications are 228 and 168 consecutively. In all the previous cases, we have discarded dc-coefficient. In case of SBT, no multiplications are required, only subtractions are needed for computing cepstral coefficients.

5. Speaker recognition experiment

5.1. Experimental framework

5.1.1. Database for experiments

Experiments are carried out for speaker verification (SV) task. In order to evaluate the performance of various class of BT based features, we have considered multiple large population speech corpora created by NIST. These are widely used in speaker recognition system evaluation. In our experiments, we have used NIST SRE 2001 and NIST SRE 2004. The database descriptions are shown in Table 1.

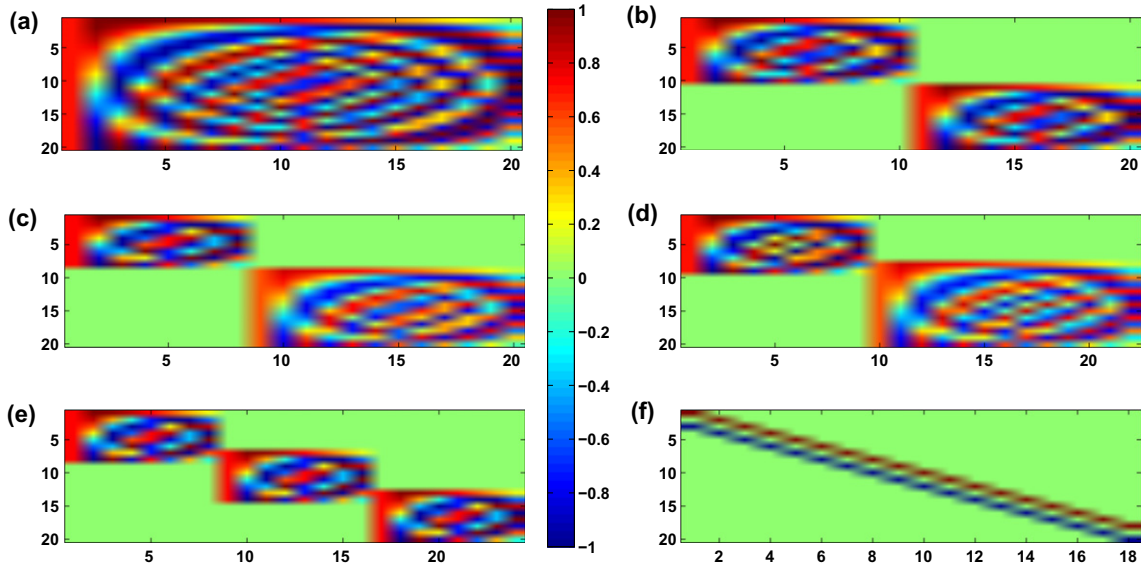

Fig. 14. Transformation kernel for different block transformations. The subfigures are shown for (a) DCT, (b) NOBT-10-10, (c) NOBT-8-12, (d) OBT-9-13, (e) OBT-8-8-8, (f) SBT. The dc-coefficients are also shown for first five types.

Apart from it, the background data for gender dependent UBM training have been collected from the development section of NIST SRE 2001 (for NIST SRE 2001 evaluation) and from NIST SRE 2003 (for NIST SRE 2004 evaluation).

5.1.2. Preprocessing

In this work, pre-processing stage is kept similar throughout different features extraction methods. It is performed using the following steps:

- The speech signal is first pre-emphasized with 0.97 pre-emphasis factor.
- The pre-emphasized speech signal is segmented into frames ($\boldsymbol{s}$) of each 20 ms, i.e. total number of samples in each frame is $N = 160$ (sampling frequency $F_s = 8$ kHz). We keep 50% overlap with adjacent frames.
- In the last step of pre-processing, each frame is windowed using hamming window.

5.1.3. Feature extraction

Standard MFCC features are extracted using linearly spaced filters in Mel scale (Kinnunen and Li, 2010). The features are further processed using RelAtive SpecTrAl (RASTA) filtering to remove the mismatch between training and testing condition. Velocity (delta) coefficients are extracted over a window of size three and those are appended with MFCC coefficient. Finally, voice activity detection (VAD) is performed to discard the non-speech frames followed by utterance level cepstral mean and variance normalization (CMVN) as a part of channel compensation and session variability reduction.

5.1.4. Speaker verification using adapted GMM

Adapted Gaussian mixture modeling based modeling technique is used to create target speaker models. The idea of GMM is to use weighted summation of multivariate Gaussian functions to represent the probability density of feature vectors as a target speaker model and it is given by,

$$p(\boldsymbol{x}) = \sum_{i=1}^{C} p_i b_i(\boldsymbol{x}), \tag{16}$$

where $\boldsymbol{x}$ is a d-dimensional feature vector, $b_i(\boldsymbol{x})$, $i = 1, 2, 3, \ldots, C$ are the component densities and p_i, $i = 1, 2, 3, \ldots, C$ are the mixture weights or *prior* of individual Gaussian.

A GMM is parameterized by the mean, covariance and mixture weights from all component densities and is denoted by

$$\lambda = \{p_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}_{i=1}^{C}. \tag{17}$$

In a speaker recognition system, each target is represented by a GMM and is referred by its model λ. The parameters of λ are optimized using iterative *expectation maximization* (EM) algorithm (Dempster et al., 1977) in *maximum likelihood* (ML) based approach. In these experiments, the GMMs are trained with a few iterations where clusters are initialized by binary splitting based vector quantization (Linde and Buzo, 1980) technique. State-of-the art speaker recognition system utilizes adapted GMM training using *maximum-a-posteriori* (MAP) approach. In this case, a large GMM of higher model order is trained as a background model, which is widely known as UBM (Reynolds et al., 2000). Individual target models are created by adapting the parameters of the UBM, i.e. mean, covariance, and priors with the training data of target speakers. It is observed that mean adaptation is sufficient to create speaker models (Kinnunen and Li, 2010).

In verification stage, the score of feature matrix of an unknown utterance $\boldsymbol{X}$ for a speaker i is determined by,

$$\theta(\boldsymbol{X}, i) = \frac{1}{T}(\log p(\boldsymbol{X}|\lambda_i) - \log p(\boldsymbol{X}|\lambda_{ubm})), \tag{18}$$

where $\boldsymbol{X}$ consists of T number of speech frames.

The scores for target and impostor trials are used to evaluate the system performance. The details of speaker verification system implementation and score calculation techniques using adapted GMM are concisely available in (Benesty et al., 2007; Kinnunen and Li, 2010).

5.1.5. Performance evaluation

Performances of SV systems are evaluated using the *detection error trade-off* (DET) plot, which is drawn with the help of DETWARE[2] tool provided by NIST. We have computed two commonly used metrics from DET curve. First, *equal error rate* (EER), the point on DET curve having equal probability of false acceptance (FA) and false rejection (FR), and second, *minimum detection cost function* (minDCF), a cost function based metric, which is computed with the same tool DETWARE by setting $C_{Miss} = 10$, $C_{False\ \ Alarm} = 1$ and $P_{Target} = 0.01$ according to the NIST evaluation plan (Przybocki and Martin, 2002; Martin and Przybocki, 2006).

5.2. Results and discussion

Speaker verification experiments are carried out on NIST SRE 2001 and 2004 corpora according to the guidelines in evaluation plan; and the experiments have been conducted on the core-test section of the databases. The pre-processing stages (such as framing, windowing, etc.) and feature post-processing schemes (like RASTA, CMVN, etc.) have been fixed for different features. We have set 20 filters in Mel filter bank, and it is fixed throughout the experiments. The different subband regions which are covered by the twenty filters are shown in Table 2. As we consider delta feature into account, the number of feature for full-band MFCC is 38. On the other hand, the feature dimension for two block NOBT and SBT

[2] http://www.itl.nist.gov/iad/mig//tools/DETware_v2.1.targz.htm.

Table 4
SV results on NIST SRE 2001 and NIST SRE 2004 using baseline MFCC (full-band DCT) and two block NOBT based MFCC feature. The lengths of the two blocks are shown in first column.

Block sizes	NIST SRE 2001			NIST SRE 2004		
	ϵ	EER (in %)	minDCF $\times$ 100	ϵ	EER (in %)	minDCF $\times$ 100
(5, 15)	0.0581	7.6546	3.4299	0.0600	14.6694	6.2178
(6, 14)	0.0552	7.8508	3.5026	0.0539	14.3696	6.1123
(7, 15)	0.0558	7.5074	3.4158	0.0549	14.1662	6.0229
(8, 12)	0.0574	7.7527	3.5063	0.0570	14.1662	5.9890
(9, 11)	0.0602	7.8999	3.5183	0.0602	14.1243	5.9883
(10, 10)	0.0605	7.7552	3.5925	0.0615	14.5436	6.0353
(11, 9)	0.0618	8.1011	3.7340	0.0626	14.5436	6.0919
(12, 8)	0.0629	7.998	3.6576	0.0632	14.9252	6.2932
(13, 7)	0.0651	8.1452	3.6385	0.0655	14.7113	6.2356
(14, 6)	0.0689	8.4396	3.7073	0.0686	15.0447	6.3397
(15, 5)	0.0710	8.5868	3.7574	0.0708	15.3004	6.3417
Full-band	0.0725	8.2434	3.5763	0.0769	14.9629	6.3231

feature will be 36. The modeling scheme has also been fixed throughout the different experiments. We have done all the SV experiments on GMM-UBM system with 256 model order. Gender independent UBMs are trained using two iterations of EM algorithm. The target models have been adapted from the UBM using relevance factor, $r = 14$ for all the cases. Top-5 Gaussians of UBM for each speech frame of test utterance have been selected for final scoring.

5.2.1. Performance of various BT based feature

The first experiment on SV is performed using two block based NOBT feature. The result for baseline MFCC and NOBT based MFCC are shown in Table 4 for different block sizes. The performances of NOBT based approaches are better than baseline MFCC in most of the cases for both the databases. The speech samples of NIST SRE 2001 are collected for matched condition only, but in the case of NIST SRE 2004, speech samples are collected by keeping considerable amount of variation in training and testing phase with diverse channels and handsets. The third column of the Table 4 shows the residual correlation measure. This metric is computed on the UBM and training set of the databases. The result shows that the block based approach has higher decorrelation power than full-band based approach. We have also found that speaker recognition performance improves if size of the first block is smaller. The reason is that if the first block is smaller then it approximately represents a formant frequency zone, F_1 and other two blocks contain two other formant frequencies, F_2 and F_3. Hence, formant frequencies or spectral peaks are independently processed. We know that F_1, F_2, and F_3 contain prominent speaker specific attributes (Quatieri, 2006). We have also observed that spectral peaks (i.e. including formant frequencies) of speech frames are concentrated in specific frequency zone. Independent processing of formant regions also provides details of peaks and valleys effectively. It is most likely one of the major reasons of improvement for block based MFCC computation. However, abrupt partitioning of filter bank energies may create trouble by ignoring full-band information. Our proposed OBT based feature overcomes this difficulty.

Table 5
SV results on NIST SRE 2001 and NIST SRE 2004 using two block OBT based feature. The specification of the two block (A and B) are shown in the first column.

Block specification	Feature dimension	NIST SRE 2001		NIST SRE 2004	
		EER (in %)	minDCF $\times$ 100	EER (in %)	minDCF $\times$ 100
A:1-9, B:8-20	40	7.2669	3.4394	13.9146	5.9420
A:1-10, B:7-20	44	7.7036	3.4319	14.0006	5.9924

Table 6
SV results on NIST SRE 2001 and NIST SRE 2004 using three block based feature. The specification of three blocks (A, B and C) are shown in first column.

Block specification	Feature dimension	NIST SRE 2001		NIST SRE 2004	
		EER (in %)	minDCF $\times$ 100	EER (in %)	minDCF $\times$ 100
A:1-8, B:9-15, C:16-20	34	8.3906	3.5466	14.5918	6.1365
A:1-8, B:7-14, C:13-20	42	7.6055	3.3892	14.3339	6.0526
A:1-8, B:9-17, C:15-20	40	7.7429	3.3986	13.9146	5.8950

We have chosen the block size (8, 12) from two block NOBT and extended this system to overlapped version. The speaker recognition performance is shown in Table 5. We observed that the performance of speaker recognition is considerably better when we keep few overlap between the adjacent blocks. However, a larger overlap keeps redundant information and the performance is degraded.

Experiments have also been also carried out using three block based system. We have considered three experiments on this tri-block feature transformation scheme. The result is shown in Table 6. The first set of result is for NOBT where the blocks are of size 8, 7, and 5. Here the performance is degraded compared to two block based NOBT technique. The reason is that we have not considered the formant regions properly. This is to note that there exits a significant amount of overlap between second and third spectral peaks as in Fig. 7. The second result is an OBT extension of the previous result where we keep an overlapping of two samples of the consecutive blocks. Hence, the formant regions are mapped to each block in a better manner compared to the earlier case. We have observed that performance is significantly improved than that of non-overlapping thee block based method. However, in the three block based approach the overall performance on two databases is degraded considerably compared to two block OBT based approach. The most probable reason is that due to the presence of noise in speech signal of SRE 2004 database the spectrum for F_2 and F_3 zone are not accurately estimated. In our next experiment, we considered the third formant (F_3) separately which is centered around 2.5 kHz within the frequency zone 1690–3300 Hz, we have precisely chosen the block size for this experiment. Among the three blocks the first block covers frequency zone related to F_1, i.e. 0–883.1663 Hz, the second block covers the zone (F_2) 745.9244–2281 Hz and the third block takes care of F_3, i.e. 1791–4000 Hz. In this case, the result is moderately improved for NIST SRE 2004 than the other cases. When speech signal is distorted by various noises, usually the critical bands in higher frequency zone are severely affected. For this reason, the performance is better when F_2 and F_3 zones are combined into a single block for both the databases. Hence, we arrive into a conclusion that the two block based OBT approach is more up to standard for the database with much variability.

We evaluated the SBT based system on both NIST SRE 2001 and 2004 databases. The DET plot of SBT along with all the other features are shown in Fig. 15. The performance of the system based on SBT is shown in Table 8. For NIST SRE 2001 database where there is no mismatch between training and testing condition performance is significantly improved over baseline MFCC and NOBT based transforms. We have got EER of 7.4583% in this case. On the other hand, the performance is degraded for NIST SRE 2004. In that case, we have obtained EER of 15.5499% using SBT compared to 14.9629% for MFCC-GMM baseline system. Most of the speech samples of NIST SRE 2004 is corrupted due the variability in telephone channel, handset, etc. As a result, most of the critical band information is distorted. The SBT feature keeps local information of frequency band by directly considering closer subband log energies, but it loses the full-band information. The full-band (or large band) information plays a significant role

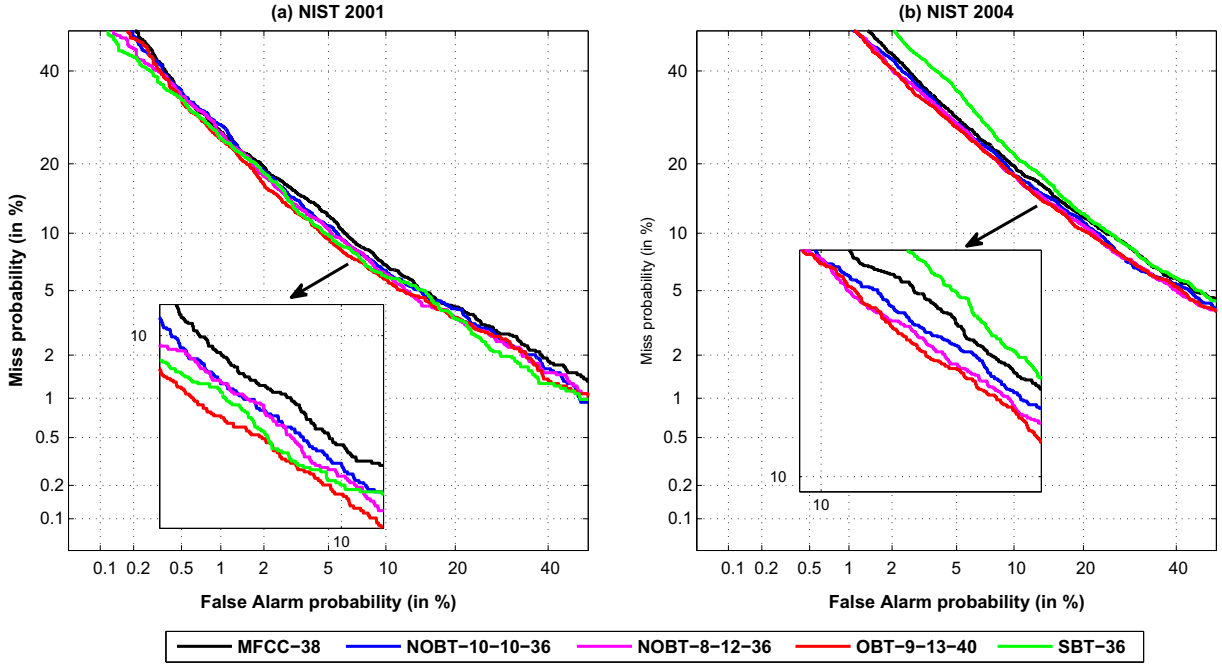

Fig. 15. Figure showing DET plot for various block based techniques. (MFCC-38: Full-band DCT based MFCC, NOBT-10-10: Double block NOBT where both blocks are of size 10, NOBT-8-12: Double block NOBT where two blocks are of sizes 8 and 12, OBT-9-13: Double block OBT where two blocks are of sizes 9 and 13, and SBT-36: SBT based feature.)

Table 7
SV results on NIST SRE 2001 and NIST SRE 2004 using PCA based feature. In PCA-40 the first coefficient is considered. In PCA-38 the first coefficient is discarded as it is equivalent to dc-coefficient.

Feature type	PCA data	NIST SRE 2001		NIST SRE 2004	
		EER (in %)	minDCF × 100	EER (in %)	minDCF × 100
PCA-40	UBM	9.4210	3.8209	18.3152	7.3806
	UBM+Train	9.0775	4.0105	18.5249	7.4448
PCA-38	UBM	8.6899	3.7433	17.0636	6.9800
	UBM+Train	8.55	3.7332	17.6863	7.2229

for recognition of speech signal degraded due to the presence of wide band noise. We have utilized the advantage of both the large band and localized information by combining their strength through score level output fusion which is discussed in Section 5.2.5.

The above study and experimental result for different block transformation based approach suggests that the two block OBT based system is a superior selection for single stream based SV experiment on NIST SRE databases. In this case, the first block approximately represents the spectral area which covers F_1 as well as the first spectral peak and other block represents the shared frequency region covered jointly by F_2 and F_3.

Table 8
SV results on NIST SRE 2001 and NIST SRE 2004 using SBT based feature.

Database	EER (in %)	minDCF × 100
NIST SRE 2001	7.4583	3.4591
NIST SRE 2004	15.5499	6.9700

5.2.2. Comparison with PCA based approach

PCA is an operation similar to KLT where the data-driven projection matrix is derived from the correlation matrix of the feature vector. PCA completely decorrelates a feature matrix, i.e. the residual correlation measure becomes exactly zero. Theoretically, PCA is the optimal decorrelation process. In this work, one of our claim is that our proposed transformation decorrelates MFLE more efficiently than DCT which helps to improve the performance. In an experiment, we have computed PCA projection matrix from MFLE of UBM data and used this for computing features. The result is shown in Table 7. It shows the result for PCA-40 and PCA-38. PCA-40 feature is the extracted using standard PCA based where

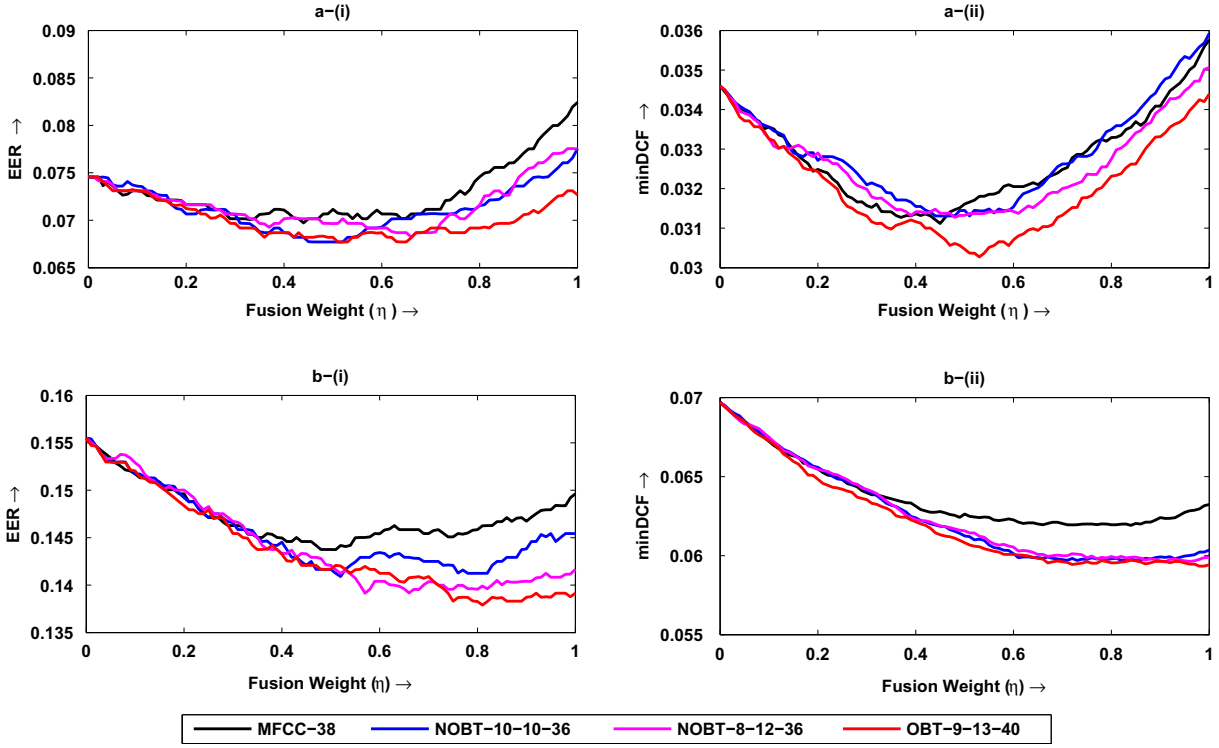

Fig. 16. Effects of fusion weight (η) on EER and minDCF are shown for (a)NIST SRE 2001 and (b)NIST SRE 2004. The variations of EER w.r.t η are shown in a-(i) and b-(i), on the other hand the changes of minDCF are shown in a-(ii) and b-(ii).

projection matrix is computed for 20 MFLE. In case of PCA-38, the first component is discarded, as it is nothing but the weighted summation of filter bank log energies. Table 7 also shows the results where the projection matrix is computed on UBM plus training data for generalizing it for target data. Clearly for all the cases the result is even far behind of standard MFCC. This is due to the data-driven approach of PCA where the projection matrix is unable to generalize the unseen test data. The result also shows that the projection matrix discarding the first coefficient is relatively better. We have also observed that the performance of PCA-38 feature where the projection matrix is computed over the combination of background and training data is comparatively better than the framework where the projection matrix is computed only on the UBM data.

5.2.3. Comparison with full covariance based approach

Full covariance based Gaussian mixture modeling is not accepted in state-of-the art speaker recognition system. The main reasons are: (a) full covariance based GMM is not robust like diagonal covariance based GMM in presence of noise and other variabilities, (b) diagonal covariance based GMM is very fast and computationally inexpensive. The determinant computation and inverse computation scheme is very straightforward for diagonal covariance based system. However, we have done one experiment for observing the performance of full covariance based GMM using full-band feature (i.e. 38 dimensional MFCC). We have obtained EER of 9.1757% and minDCF of 0.039381 for NIST SRE 2001. On the other hand, we have got EER of 17.854% and minDCF of 0.074813 for NIST SRE 2004. Though this method considers the off-diagonal elements of covariance matrix, still the robustness of the system degrades severely due to the consideration of the data. Note that the CPU time for training and testing is also increased by several factors for full covariance based method.

Table 9
SV results on NIST SRE 2001 and NIST SRE 2004 for different fused system where baseline spectral feature based systems are fused with SBT based system. Fusion weight (η) is set at 0.5 for NIST 2001 and 0.8 for NIST 2004.

Baseline feature	NIST SRE 2001		NIST SRE 2004	
	EER (in %)	minDCF × 100	EER (in %)	minDCF × 100
MFCC-38	7.1148	3.1611	14.5855	6.1953
NOBT-10-10-36	6.7713	3.1450	14.1243	5.9798
NOBT-8-12-36	6.9652	3.1324	13.9566	5.9944
OBT-9-13-40	6.8204	3.0468	13.8266	5.9545

5.2.4. Dimensionality issue

Feature dimension is an important issue in designing speaker recognition system for realtime applications. The dimension of our proposed features based on NOBT and SBT are lesser compared to standard MFCC based features. In case of NOBT, as we usually discard the dc-coefficients, the feature dimension reduces significantly with the increase in number of blocks. However, the number of features for OBT increases linearly with the increase in number of blocks and overlap. Hence, we have restricted the feature dimension by keeping block size to two for OBT with an overlap of 2 subbands. Therefore, the dimension of our proposed feature (i.e. OBT-9-13) becomes 40 which is slightly larger compared to 38 of standard MFCC extracted using 20 filters. It is shown that 40 dimensional

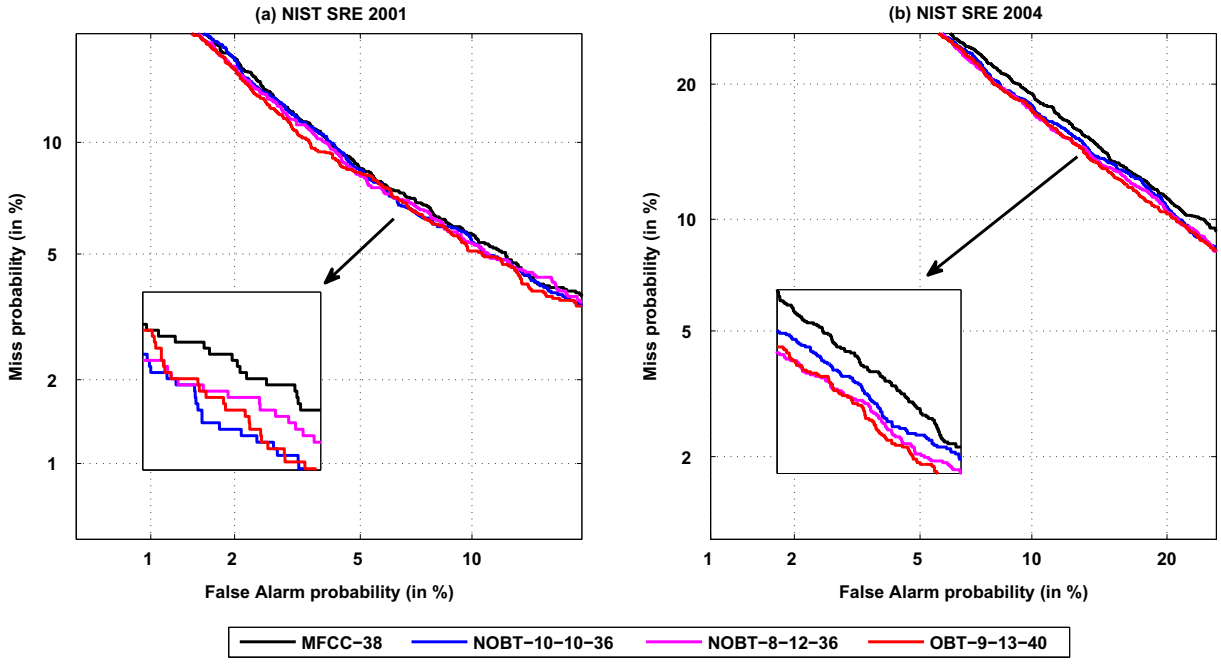

Fig. 17. Figure showing DET plot for different fused system. The fusion weights (η) are 0.5 (for NIST SRE 2001) and 0.8 (for NIST SRE 2004).

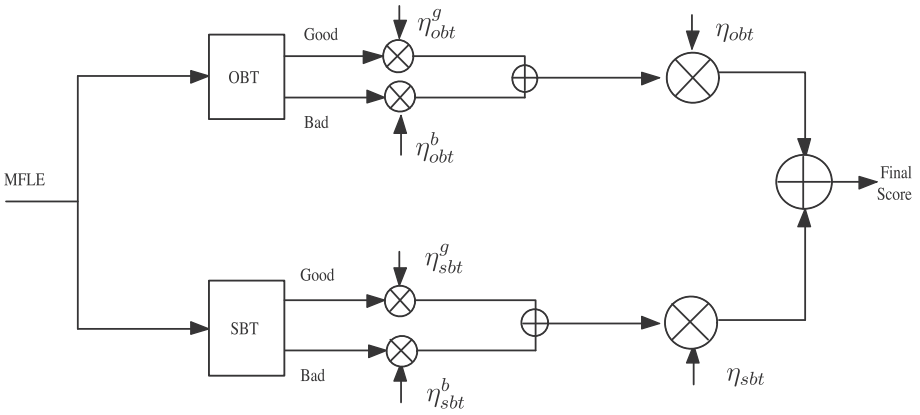

Fig. 18. Block diagram of the proposed missing feature based testing scheme. Good and Bad denote log likelihood ratio of reliable features and unreliable features correspondingly.

feature vector computed using two overlapped blocks of size 9 and 13 are reasonably better than other multi-block cases. We have also compared the performance of this 40 dimensional proposed feature with 40 dimensional standard MFCC feature extracted using 21 filters. We have found EER of 8.1992% and minDCF of 0.037080 for NIST SRE 2001. On the other hand, for NIST SRE 2004, we have obtained EER of 15.0887% and minDCF of 0.063171. The performance of our proposed feature remains consistently better for both the databases.

Table 10
SV results on NIST SRE 2001 in the presence of different types of additive noises. The noise samples are taken from NOISEX-92 database. The results are shown for different SNRs (20 dB, 10 dB, and 0 dB) for single stream based system.

SNR	Feature		Noise type			
			White	Hfchannel	Babble	Pink
20 dB	MFCC-38	EER (in %)	13.3955	13.2974	8.4396	10.1079
		minDCF × 100	6.0316	5.8236	3.7954	4.3091
	NOBT-10-10-36	EER (in %)	12.1197	12.8067	7.9612	9.0775
		minDCF × 100	5.4858	5.6605	3.7021	4.1943
	NOBT-8-12-36	EER (in %)	12.1246	12.3135	8.0864	9.0775
		minDCF × 100	5.4578	5.5303	3.5563	4.0368
	OBT-9-13-40	EER (in %)	12.0707	12.5123	7.7453	8.5942
		minDCF × 100	5.4710	5.6982	3.5118	3.9993
	SBT-36	EER (in %)	12.9588	15.1129	8.3415	8.8445
		minDCF × 100	5.9345	7.0027	3.7332	4.1056
10 dB	MFCC-38	EER (in %)	20.6575	21.1065	13.5427	17.4681
		minDCF × 100	9.0520	8.8470	6.6280	8.2390
	NOBT-10-10-36	EER (in %)	18.9941	20.5643	13.1501	17.0265
		minDCF × 100	8.2653	8.4828	6.1758	7.6323
	NOBT-8-12-36	EER (in %)	20.0613	20.7115	13.3464	17.0216
		minDCF × 100	8.4405	8.4209	6.1341	7.5782
	OBT-9-13-40	EER (in %)	20.5103	21.7861	13.5427	17.5687
		minDCF × 100	8.6945	8.6612	6.2836	8.0261
	SBT-36	EER (in %)	22.7208	27.3307	15.9470	20.0196
		minDCF × 100	9.2348	9.8256	7.6837	8.6481
0 dB	MFCC-38	EER (in %)	33.8543	32.5368	29.8332	30.8734
		minDCF × 100	10.0000	10.0000	9.8766	10.0000
	NOBT-10-10-36	EER (in %)	31.5015	31.7468	28.9990	29.1462
		minDCF × 100	9.9754	9.9410	9.7929	9.9606
	NOBT-8-12-36	EER (in %)	32.9735	32.9244	28.6114	28.9622
		minDCF × 100	9.9801	9.9509	9.8604	9.9647
	OBT-9-13-40	EER (in %)	33.4151	34.2493	29.5388	30.6183
		minDCF × 100	9.9802	9.9656	9.8064	9.9508
	SBT-36	EER (in %)	37.1369	39.2051	33.9647	33.8567
		minDCF × 100	9.9901	9.9803	9.9340	9.9652

5.2.5. Performance of fused system

The performance of the speaker recognition is further improved with help of score level output fusion. The

Table 11
SV results on NIST SRE 2001 in the presence of different types of additive noises. The noise samples are taken from NOISEX-92 database. The results are shown for different SNRs (20 dB, 10 dB, and 0 dB) for fused system. The fusion weight is set at 0.6 for 20 dB SNR, 0.7 for 10 dB SNR, and 0.9 for 0 dB SNR.

SNR	Feature		Noise type			
			White	Hfchannel	Babble	Pink
20 dB	MFCC-38	EER (in %)	12.2669	13.1477	7.6055	8.2434
		minDCF × 100	5.4166	5.8592	3.3696	3.7868
	NOBT-10-10-36	EER (in %)	11.5309	12.7576	7.1148	7.8974
		minDCF × 100	5.1449	5.8862	3.2341	3.6923
	NOBT-8-12-36	EER (in %)	11.5800	12.7576	7.1271	8.2434
		minDCF × 100	5.1677	5.8083	3.2026	3.6203
	OBT-9-13-40	EER (in %)	11.7738	13.0520	7.1639	7.7036
		minDCF × 100	5.2470	5.9110	3.2427	3.6350
10 dB	MFCC-38	EER (in %)	20.1178	22.0314	13.0029	17.0731
		minDCF × 100	8.6257	8.8213	6.3898	7.9532
	NOBT-10-10-36	EER (in %)	18.7414	21.4843	12.9171	16.3395
		minDCF × 100	8.1694	8.7847	6.1947	7.5290
	NOBT-8-12-36	EER (in %)	19.6320	21.8842	13.0029	16.6830
		minDCF × 100	8.3371	8.7854	6.2185	7.6302
	OBT-9-13-40	EER (in %)	20.1668	22.5711	13.4446	17.0805
		minDCF × 100	8.5169	8.9414	6.3286	7.8816
0 dB	MFCC-38	EER (in %)	33.7120	32.8263	30.0736	30.6673
		minDCF × 100	10.0000	10.0000	9.8861	10.0000
	NOBT-10-10-36	EER (in %)	31.7959	31.8940	29.3916	29.0481
		minDCF × 100	9.9754	9.9508	9.8273	9.9605
	NOBT-8-12-36	EER (in %)	33.1158	33.3783	28.7022	29.1462
		minDCF × 100	9.9801	9.9558	9.8704	9.9598
	OBT-9-13-40	EER (in %)	33.9058	34.7301	29.4406	30.4809
		minDCF × 100	9.9753	9.9605	9.8365	9.9508

system combination scheme, which is employed here, is computationally less expensive compared to feature level fusion for GMM based speaker recognition system. We adopted the simplest fusion strategy: *linear fusion*. The score, i.e. log-likelihood ratio for two sub-systems are weighted and summed to derive the final score. If the log-likelihood score of two systems X and Y are LLR_x and LLR_y, then final score is $LLR_{fused} = \eta \cdot LLR_x + (1-\eta) \cdot LLR_y$, where η is fusion weight and $0 < \eta < 1$. The speaker recognition performance for fused system is shown in Fig. 16 for different values of η. Various large band information based system is fused with SBT based system with a fixed fusion weight η. As the speech quality of NIST SRE 2001 is considerably good, we have chosen fusion weight of 0.5, i.e. equal importance is given to both the systems. On the other hand, we have empirically chosen fusion weight 0.8 for large band information in the evaluation of NIST SRE 2004 database. This is due to the fact that SBT information are less reliable for distorted speech signals. The DET plot for combined system is shown in Fig. 17 for both the databases. The plot depicts that the fused system based on the combination of OBT-9-13 and SBT are better than the other systems in terms of EER. In Table 9, the speaker recognition result is shown for fusion of SBT with various other features. The combined performance is improved for single stream based system for all the cases. Particularly, for block based feature extraction, the performance improvement is relatively significant. The performance improvement is observed for both EER and minDCF. We obtained an EER of 6.8204% for NIST 2001 database when OBT based system with two sample overlapping (i.e. OBT-9-13 based system) is fused with SBT based system. At the same time, we obtained relative minDCF improvement of 14.81% over baseline MFCC. Though for this particular database the EER is not much reduced for fused scheme based on OBT-9-13 feature compared to other block based techniques, the minDCF is significantly improved over other cases. Note that minDCF plays a significant role in computing verification threshold for real time speaker recognition system. On the other hand, in NIST SRE 2004 also, we have obtained best performance for the case where OBT-9-13 is combined with SBT. There we have achieved EER of 13.8266% and minDCF of 5.9546%. In that case, the relative improvement over baseline MFCC is 7.59% in EER and 5.83% in minDCF.

5.2.6. Performance in presence of noise

The performance of speaker recognition systems based on proposed BTs are evaluated on noisy speech data. We

Table 12
SV results on NIST SRE 2001 in the presence various of narrow-band noises. The results are shown for different SNRs (20 dB, 10 dB, and 0 dB) for fused system. The fusion weight is set at 0.6 for 20 dB SNR, 0.7 for 10 dB SNR, and 0.9 for 0 dB SNR.

SNR	Feature		Single stream		Fused with SBT	
			Type-I	Type-II	Type-I	Type-II
20 dB	MFCC-38	EER (in %)	16.2929	15.265	15.4073	15.3091
		minDCF × 100	6.4260	6.4967	6.2508	6.6075
	NOBT-10-10-36	EER (in %)	14.8184	14.8184	14.3768	15.3091
		minDCF × 100	5.8941	6.3276	5.9183	6.4382
	NOBT-8-12-36	EER (in %)	16.7812	14.7203	15.7998	15.3680
		minDCF × 100	6.4745	6.2480	6.2304	6.4500
	OBT-9-13-40	EER (in %)	15.2601	14.6222	15.1178	15.5986
		minDCF × 100	6.0955	6.3139	6.0878	6.5292
	SBT-36	EER (in %)	17.4190	19.5780	-	-
		minDCF × 100	7.0402	8.0275	-	-
10 dB	MFCC-38	EER (in %)	22.8656	23.1600	22.4730	24.3376
		minDCF × 100	8.4021	8.8851	8.3266	8.9446
	NOBT-10-10-36	EER (in %)	20.4612	23.1992	20.3140	23.7488
		minDCF × 100	7.7507	8.6676	7.7759	8.8152
	NOBT-8-12-36	EER (in %)	23.9450	23.5034	22.8165	23.9450
		minDCF × 100	8.6625	8.9654	8.4552	8.9912
	OBT-9-13-40	EER (in %)	22.1320	24.7301	21.9431	25.0245
		minDCF × 100	8.1286	8.8062	8.1646	8.9903
	SBT-36	EER (in %)	24.5437	30.0834	-	-
		minDCF × 100	8.8821	9.9266	-	-
0 dB	MFCC-38	EER (in %)	26.3984	28.4985	26.1482	28.8027
		minDCF × 100	9.3465	9.8261	9.3051	9.8210
	NOBT-10-10-36	EER (in %)	24.7301	30.8072	24.3817	31.0059
		minDCF × 100	8.9592	9.6588	8.9353	9.6794
	NOBT-8-12-36	EER (in %)	27.1860	30.7139	26.8916	30.2380
		minDCF × 100	9.5948	9.8476	9.4942	9.8632
	OBT-9-13-40	EER (in %)	25.9593	32.7355	25.8096	32.5736
		minDCF × 100	9.0763	9.9077	9.0719	9.9595
	SBT-36	EER (in %)	28.8494	37.5442	-	-
		minDCF × 100	9.5892	9.9951	-	-

have observed the effect for both standard noise and synthetic narrow-band noise. The standard noise database NOISEX-92 is used for the evaluation. The noise samples are first down sampled at 8 kHz, then adaptively scaled to a desired SNR value and added to the required speech signal. We have chosen NIST SRE 2001 corpus only for this particular experiment. NIST SRE 2004 is not selected because it is already distorted by various channel noise and handset effects. The experiments have been conducted using four types of standard additive noise (white, hfchannel, babble and pink) for different levels of SNRs (20 dB, 10 dB and 0 dB). The results for single stream based system are shown in Table 10. We can observe that the performance of proposed BT based systems are better than DCT based system in most of the cases. The speaker recognition performance can be further improved with fused system. Score of full-band based feature as well as proposed large block based BT techniques are fused together with SBT feature based scores. As the performance of SBT becomes very poor with increase in noise we have adjusted the fusion weight empirically for different cases. We have kept fusion weight 0.6 for 20 dB noise, 0.7 for 10 dB noise and 0.9 for 0 dB noise. The result for dual stream based system is shown in Table 11. Here furthermore we observe that the performance of fused system is significantly better than that of single stream based system for almost all kinds of noise of various SNRs.

The block based features are affected with different degree in presence of narrow-band noise. In order to observe this effect, we conducted experiments on narrow-band noise. Speaker recognition results using two types of narrow-band signal are shown in Table 12. In left half of the table, the results are shown for the single stream based system and the performance of the fused system is shown in right half of the table. Though the performance of block based features are better for some of the cases, it is not better consistently. This is partially due to the energy based VAD used in our experiments which 'wrongly' treats the non-speech higher energy based frames as speech frames. In order to show the effect, we have computed the block based features for a complete utterance and taken the average over all the frames. Now to study the effect of voice activity detection, we have done this experiment twice. In the first case, all the frames of noisy speech, which are declared as speech by energy based VAD, are considered and the result is shown in Fig. 19. We can see that features which are not extracted from noisy zone are also affected. The reason is that originally non-speech frames, which are detected as speech in presence of noise, create this error. We have confirmed this after doing experiment using the speech VAD labels extracted from the clean speech signal and applying it to noisy data. The result is shown in Fig. 20 and it is clear that the narrow-band noise has only local effect.

However, the advantages of block based feature can be retrieved using MFT technique. In this method, the contribution from different features are operated separately. In multidimensional statistical methods, scores of unreliable features are either masked or given lesser importance by weighting with a lesser value than that of reliable features. As in block based feature extraction scheme, different blocks are unequally affected due to narrow-band noise,

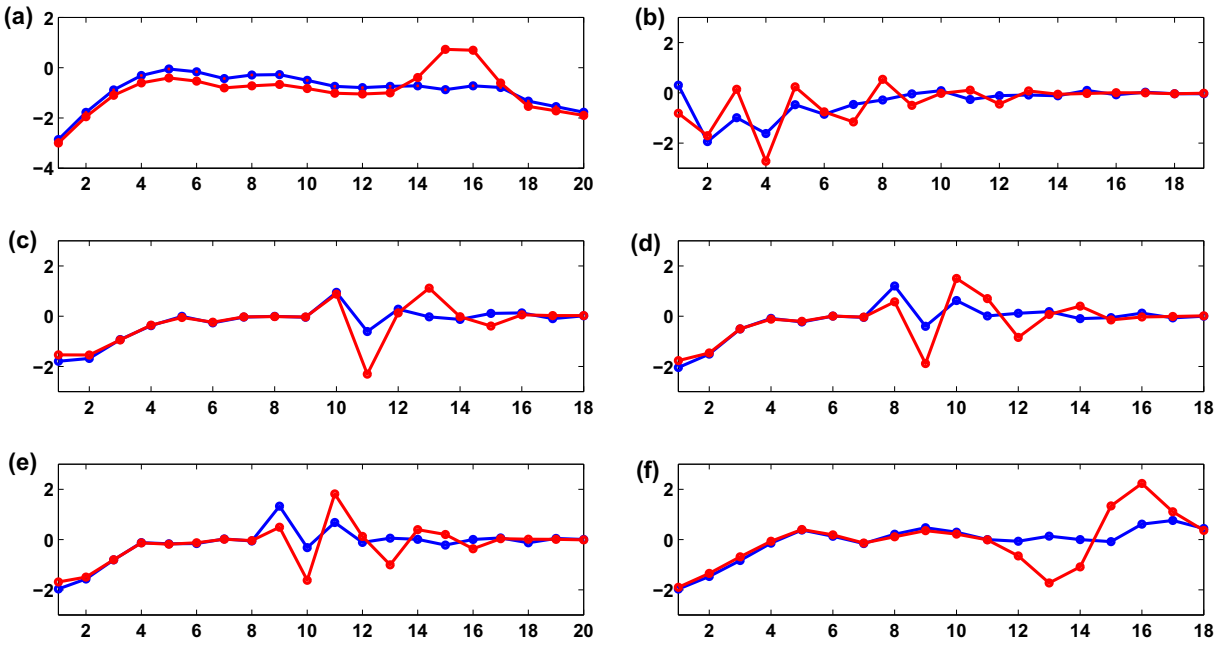

Fig. 19. The effect of narrow-band noise (SNR:15 dB) on (a) mel filter bank log energy and on different features based on (b) DCT, (c) NOBT-10-10, (d) NOBT-8-12, (e) OBT-9-13, and (f) SBT (Blue line: clean speech, red line: noisy speech.) is shown for a complete speech utterance. Average of the parameters (e.g. MFLE, features) are computed over all the voiced frames of the *noisy* utterance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

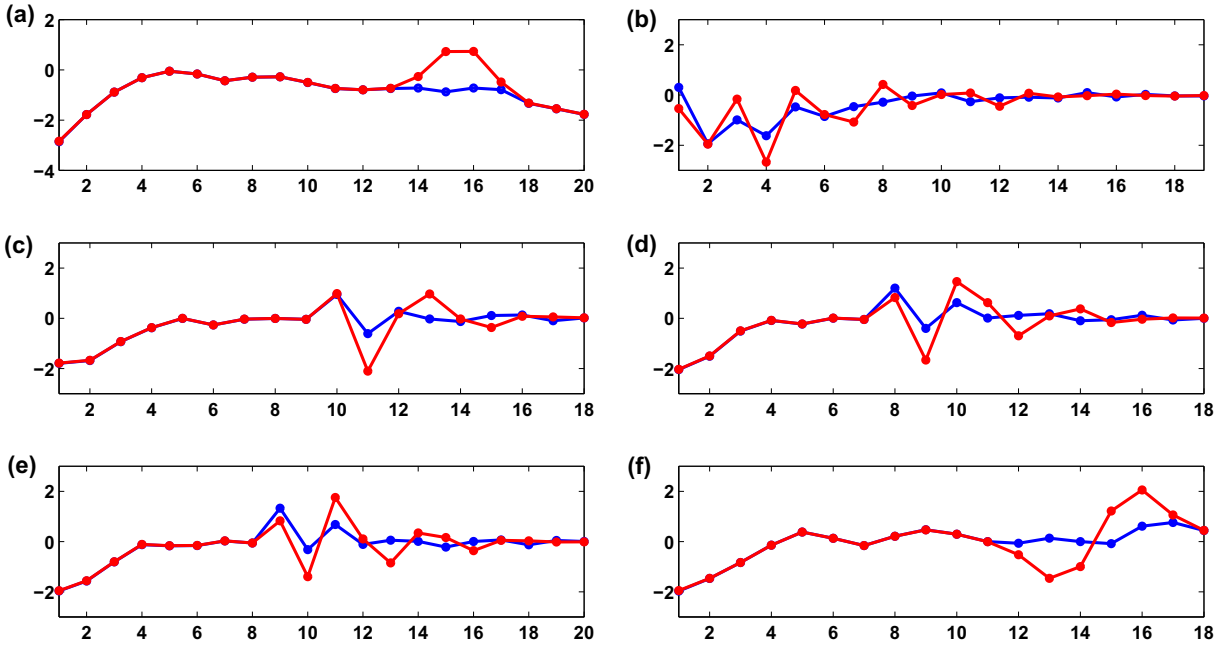

Fig. 20. The effect of narrow-band noise (SNR:15 dB) on (a) mel filter bank log energy and on different features based on (b) DCT, (c) NOBT-10-10, (d) NOBT-8-12, (e) OBT-9-13, and (f) SBT (Blue line: clean speech, red line: noisy speech.) is shown for a complete speech utterance. Average of the parameters (e.g. MFLE, features) are computed over all the voiced frames of the corresponding *clean speech* utterance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

we have treated them separately. We have chosen the OBT-9-13 feature and applied missing feature theory in fused mode. Let us assume that we have prior information concerning the nature of narrow-band noise. Both kinds of MFT technique are applied. In masking, as the second block is much affected with noise we only compute the score for first block, i.e. first 8 dimensions including the corresponding deltas are selected for this block. Hence, the feature dimension becomes 16. On the other hand, we have chosen first 11 and 18th SBT based feature and corresponding deltas, i.e. 22 dimensions from 36 dimensional SBT feature. The previous fusion weights, i.e. 0.6, 0.7 and 0.9 are chosen for 20 dB, 10 dB and 0 dB. The result for masking based MFT technique is shown in Table 13. Incidentally, the unreliable features are not completely ineffective for speaker recognition. They also have less but not negligible contribution depending on the SNR of the signal. Therefore, the result can be further improved if we consider the contribution from the distorted blocks. This scheme is shown in Fig. 18. Here, the contribution from both *good* and *bad* features are considered. As the bad features are more distorted with the increase in noise we keep higher weight to the good feature scores than the bad feature scores. The contribution of the good feature is linearly increased with the increase in noise power. We have chosen η^g_{obt}, η^g_{sbt}, and η_{obt} identical for a particular SNR. We have set them at 0.6, 0.7, and 0.9 for 20 dB, 10 dB, and 0 dB correspondingly, i.e. same weighting as of fusion scheme for noisy data. It has been empirically found that these are reasonably better than any other arbitrary weighting. The results for weighting based MFT scheme are shown in right half of the Table 13. We have acquired significant performance improvement in low SNR for both types of narrow-band noise. This scheme could be further improved with automatic detection of narrow-band noise and effective frame based weight selection.

Table 13
SV results in the presence of narrow-band noise on NIST SRE 2001 using missing feature theory. Overlapped block transform based fused scheme's result is shown.

SNR	Feature	Masking		Weighting	
		Type-I	Type-II	Type-I	Type-II
20 dB	EER (in %)	15.1717	17.3700	14.6099	16.0942
	minDCF × 100	6.7106	7.2827	6.1362	6.5296
10 dB	EER (in %)	19.5780	22.1295	19.1855	22.1835
	minDCF × 100	8.1150	8.9990	7.8261	8.8185
0 dB	EER (in %)	25.1227	27.1467	20.9151	22.3258
	minDCF × 100	9.4888	9.7222	8.6014	9.0407

6. Conclusions

Block based MFCC computation schemes are efficient and robust in speaker recognition context. In this paper, we have investigated an improved block based approach for speaker recognition. First, the feature extraction schemes using non-overlapped and overlap block transformation are analytically formulated. Proposed block transform fairly decorrelates filter bank log energies as an alternative of standard DCT. The experimental evaluation is performed on standard databases, and this shows that formant specific block transformations perform better.

We also propose a novel block based orthogonal transformation technique which captures transitional information of log-energies of filter bank in frequency axis. The information covered in the later approach contains complementary attribute to former block transforms. The performance of speaker recognition system is further enhanced by combining the strength of both kinds of features using score level linear fusion. We have obtained substantial performance improvement for both standard performance evaluations metric, i.e. EER and minDCF. The proposed system is very much suitable for speaker recognition in noisy condition specifically for narrow-band noise. In our current work, we have mostly focused on the effectiveness of block based linear transformation for improving the performance. The performance could be further enhanced by effective processing of the signal in subband level, i.e. subband filtering, non-linear operations, etc. As a final point in current work we have used very basic fusion scheme which could be replaced with advanced logistic regression based fusion technique for better system design. An investigation can be carried out in frame level score combination for more effective speaker recognition system development. We can summarize that our methodical study on block transform and its evaluation in NIST SRE databases could be a new groundwork for feature level development of modern speaker recognition system.

Acknowledgement

The authors would like to thank anonymous reviewers for their useful comments that helped revising the paper.

References

Ahmed, N., Natarajan, T., Rao, K., 1974. Discrete cosine transfom. IEEE Trans. Comput. C-23 (1), 90–93.

Akansu, A.N., Haddad, R.A., 1992. Multiresolution signal decomposition: Transforms, subbands, and wavelets. Academic Press.

Benesty, J., Sondhi, M., Huang, Y., 2007. Springer Handbook of Speech Processing, first ed. Springer-Verlag, Secaucus, NJ.

Besacier, L., Bonastre, J.-F., 2000. Subband architecture for automatic speaker recognition. Signal Process. 80 (7), 1245–1259.

Besacier, L., Jean-Frantois, B., 1997. Subband approach for automatic speaker recognition: Optimal division of the frequency domain. Lecture Notes Comput. Sci. 1206, 193–202.

Campbell, J., Shen, W., Campbell, W., Schwartz, R., Bonastre, J.-F., Matrouf, D., 2009. Forensic speaker recognition. IEEE Signal Process. Mag. 26 (2), 95–103.

Campbell, J.P., Jr, 1997. Speaker recognition: A tutorial. Proc. IEEE 85 (9), 1437–1462.

Chakroborty, S., 2008. Some studies on acoustic feature extraction, feature selection and multi-level fusion strategies for robust text-independent speaker identification. Ph.D. thesis, Indian Institute of Technology Kharagpur.

Chakroborty, S., Saha, G., 2010. Feature selection using singular value decomposition and qr factorization with column pivoting for text-independent speaker identification. Speech Comm. 52 (9), 693–709.

Chetouani, M., Faundez-Zanuy, M., Gas, B., Zarader, J., 2009. Investigation on lp-residual representations for speaker identification. Pattern Recognition 42 (3), 487–494.

Damper, R.I., Higgins, J.E., 2003. Improving speaker identification in noise by subband processing and decision fusion. Pattern Recognition Lett. 24 (13), 2167–2173.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Statist. Soc. Ser. B (Methodol.) 39, 1–38.

Douglas, O., 2009. Speech Communications: Human and Machine, second ed. Universities Press.

Finan, R., Damper, R., Sapeluk, A., 2001. Improved data modeling for text-dependent speaker recognition using sub-band processing. Internat. J. Speech Technol. 4 (1), 45–62.

Garreton, C., Yoma, N., Torres, M., 2010. Channel robust feature transformation based on filter-bank energy filtering. IEEE Trans. Audio Speech Lang. Process. 18 (5), 1082–1086.

Hung, W.-W., Wang, H.-C., 2001. On the use of weighted filter bank analysis for the derivation of robust mfccs. IEEE Signal Process. Lett. 8 (3), 70–73.

Jain, A.K., 2010. Fundamentals Of Digital Image Processing, first ed. PHI Learning Pvt. Ltd.

Jingdong, C., Paliwal, K., Nakamura, S., 2000. A block cosine transform and its application in speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing (INTERSPEECH 2000 – ICSLP), Vol. IV, pp. 117–120.

Jingdong, C., Yiteng, H., Qi, L., Paliwal, K., 2004. Recognition of noisy speech using dynamic spectral subband centroids. IEEE Signal Processing Lett. 11 (2), 258–261.

Jung, S.-H., Mitra, S., Mukherjee, D., 1996. Subband dct: Definition, analysis, and applications. IEEE Trans. Circ. Syst. Video Technol. 6 (3), 273–286.

Kajarekar, S., Yegnanarayana, B., Hermansky, H., 2001. A study of two dimensional linear discriminants for asr. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, 2001 (ICASSP), 2001, Vol. 1, pp. 137–140.

Kim, S., Ji, M., Kim, H., 2008. Noise-robust speaker recognition using subband likelihoods and reliable-feature selection. ETRI J. 30 (1), 89–100.

Kinnunen, T., 2004. Spectral features for automatic textindependent speaker recognition. Ph.D. thesis, University of Joensuu.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Comm. 52 (1), 12–40.

Kwon, O., Lee, T., 2004. Phoneme recognition using ica-based feature extraction and transformation. Signal Process. 84, 1005–1019.

Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantization design. IEEE Trans. Comm. COM-28 (4), 84–95.

Lippmann, R., Carlson, B., 1997. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise. EUROSPEECH, KN37–KN40.

Mak, B., 2002. A mathematical relationship between full-band and multiband mel-frequency cepstral coefficients. IEEE Signal Process. Lett. 9 (8), 241–244.

Malvar, H., Staelin, D., 1989. The lot: Transform coding without blocking effects. IEEE Trans. Acous. Speech Signal Process. 37 (4), 553–559.

Martin, A., Przybocki, M., 2006. 2004 nist speaker recognition evaluation. Linguistic Data Consortium.

Ming, J., Hazen, T., Glass, J., Reynolds, D., 2007. Robust speaker recognition in noisy conditions. IEEE Trans. Audio Speech Lang. Process. 15 (5), 1711–1723.

Mukherjee, J., Mitra, S., 2002. Image resizing in the compressed domain using subband dct. IEEE Trans. Circ. Syst. Video Technol. 12 (7), 620–627.

Nasersharif, B., Akbari, A., 2007. Snr-dependent compression of enhanced mel sub-band energies for compensation of noise effects on mfcc features. Pattern Recognition Lett. 28, 1320–1326.

Nitta, T., Takigawa, M., Fukuda, T., 2000. A novel feature extraction using multiple acoustic feature planes for hmm-based speech recognition. ICSLP 1, 385–388.

Oppenheim, A.V., Schafer, R.W., 1979. Digital Signal Processing. Prentice-Hall, Inc.

Przybocki, M., Martin, A., 2002. 2001 nist speaker recognition evaluation corpus. Linguistic Data Consortium.

Quatieri, T., 2006. Discrete-time Speech Signal Processing. Prentice-Hall, Upper Saddle River, NJ.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Process. 10 (1–3), 19–41.

Sahidullah, M., Chakroborty, S., Saha, G., 2010. On the use of perceptual line spectral pairs frequencies and higher order residual moments for speaker identification. Internat. J. Biomet. 2, 358–378.

Sahidullah, M., Saha, G., 2009. On the use of distributed dct in speaker identification. In: India Conference (INDICON), 2009 Annual IEEE, pp. 1 –4.

Sivakumaran, P., Ariyaeeinia, A.M., Loomes, M., 2003. Sub-band based text-dependent speaker verification. Speech Comm. 41 (2–3), 485–509.

Takiguchi, T., Ariki, Y., 2007. Pca-based speech enhancement for distorted speech recognition. J. Multimedia 2 (5), 13–18.

Vale, E., Alcaim, A., 2008. Adaptive weighting of subband-classifier responses for robust text-independent speaker recognition. Electron. Lett. 44 (21), 1280–1282.