A BLOCK COSINE TRANSFORM AND ITS APPLICATION IN SPEECH RECOGNITION

Jingdong Chen †*, Kuldip K. Paliwal † and Satoshi Nakamura *

† School of Microelectronic Engineering, Griffith University
Brisbane, QLD 4111, Australia
* ATR Spoken Language Translation Research Labs
Kyoto, 619-0288, Japan
E-mail: {jingdong.chen, nakamura}@slt.atr.co.jp, k.paliwal@me.gu.edu.au

ABSTRACT

Noise robust speech recognition has become an important area of research in recent years. The fact that human listeners can recognize speech in the presence of strong noise inspires researchers to imitate some aspects of human auditory perception in automatic speech recognition. This has led to sub-band based speech recognition in which the full-band speech is split into several sub-bands and where each sub-band is processed separately. The resulting multi-band features can be combined in various ways for carrying out speech recognition task. Reported results have shown the superiority of this technique for speech recognition in strong noise conditions. In this paper, we will briefly review the multi-band feature extraction. We will then propose a block discrete cosine transform (BDCT) with its kernel transformation matrix being derived from the decomposition of the kernel of the discrete cosine transform (DCT). We show that the BDCT approximates the DCT in keeping information in decorrelating a sequence. When the BDCT is applied to the mel frequency filter bank energies (FBEs) to replace the DCT to convert them to cepstral coefficients, a new kind of MFCCs is yielded. We call these new features Block discrete cosine transform based MFCCs (BMFCCs) and show that a sub-band processing idea is implicit in the BMFCCs since the BDCT automatically divides the mel frequency FBEs into two sub-bands. We will report various speech recognition results using the BMFCCs as well as the comparison with the multi-band MFCCs and full-band MFCCs to elaborate the properties of the BMFCCs.

1. INTRODUCTION

Significant advances have been made in recent years in the area of automatic speech recognition. It is now possible to use a speech recognition successfully in a controlled environment. However, the performance of a speech recognizer suffers dramatic degradation when there is a mismatch between training and testing environments [1-3]. There are many factors that contribute to this mismatch. The main factor, that causes the mismatch, is the presence of ambient background noise in the speech signal. Maintaining good recognition accuracy in noisy conditions has become one of the challenging areas of research currently.

The fact that human listeners can achieve and secure very high recognition accuracy even in the conditions in which the signal to noise ratio becomes extremely low inspires researchers to mimic some aspects of human auditory perception in automatic speech recognition. Psychoacoustic evidence shows that human beings process speech on a narrow band basis. An intuitive way to imitate the auditory system is to split the full-band speech into several sub-bands and represent each sub-band individually. This has led to a technique called sub-band based speech recognition.

One straightforward way to achieve sub-band representation is to divide the full-band speech signal into several sub-bands and convert each sub-band spectrum into several cepstral features. These sub-band features are then concatenated together as a single feature vector and used for speech recognition [4-7,12-14]. Results reported in the literature have shown the advantages of these multi-band features for noisy speech recognition. However, the performance for clean speech is often poorer as the features from different sub-bands may be correlated. Furthermore, the number of sub-bands and the boundaries for each sub-band are empirical values which have to be manually adjusted to gain good performance for a given recognition task.

An alternative is to model the sub-band features independently and to combine the likelihood score at some segmental level [8-10]. Such a combination may enable a more flexible way to manipulate the sub-band features to permit further enhancement in performance. An unsolved problem with this approach is how to determine the weighting function to guarantee at least a sub-optimal combination of sub-band features.

In this paper, we will firstly review the extraction of multi-band MFCCs. We will then propose a block discrete cosine transform (BDCT) with its kernel transformation matrix being derived from the decomposition of the kernel of the discrete cosine transform (DCT). We show that the proposed new transform behaves similarly with the DCT in keeping information in decorrelating a sequence.

When the BDCT is applied to the representation of the mel frequency cepstrum in replacing the DCT, a new type of MFCCs is obtained. We call these new MFCCs Block discrete cosine transform based MFCCs (BMFCC). It is found that the BDCT automatically divide the power spectrum into two sub-bands, hence a sub-band processing idea is implicit in the new cepstral features.

Various speech recognition experiments are carried out to test the properties of the BDCT and the BMFCCs. We will report some results as well the comparison with the multi-band and full-band MFCCs to elaborate the properties of these new features.

2. MULTI-BAND MFCC

Let $E = [e_1, e_2, \cdots, e_N]$ denote a sequence of log filter bank energies, where N is the number of filters in the filter bank, then the full-band cepstrum is computed from a DCT,

$$X_F = [C]E \tag{1}$$

Suppose we divide the speech signal into M sub-bands. In order to compute cepstral coefficients for the *m*th sub-band, we process each sub-band signal with a filter bank having N_m filters. Thus these filter bank energies are given as

$E=[E_1,\ E_2,\ \cdots,\ E_M]=[\{e_{11},e_{12},\cdots,e_{1N_1}\},\ \{e_{21},e_{22},\cdots,e_{2N_2}\},\ \cdots,\ \{e_{M1},e_{M2},\cdots,e_{MN_M}\}]$. The cepstral coefficients for the *m*th sub-band can be computed through a DCT as follows:

$$X_S^m=[C]_m E_m \tag{2}$$

Various ways can be used to merge the sub-band cepstral vectors. If these vectors are directly concatenated together as a single feature vector, i.e., $X_S=[X_S^1,X_S^2,\cdots,X_S^M]$, this merging strategy is called feature combination (FB) [5] and the resulting cepstra are called multi-band features.

3. BLOCK DISCRETE COSINE TRANSFORM

The DCT, which has been found to be asymptotically equivalent to the optimal Karhunen-Loeve transform (KLT) in decorrelating a signal sequence, has shown its special applicability in cepstral analysis of speech signal. For a vector X containing N sample data points, its DCT is a vector X_c given by

$$X_c=[C]X \tag{5}$$

where $[C]$ is a N by N kernel conversion matrix of the DCT. The *(m,n)* element of this matrix is defined as [15]

$$C_N^{m,n}=\left(\frac{2}{N}\right)^{1/2}\left[k_m\cos\left(\frac{m(n+0.5)\pi}{N}\right)\right]\quad m,n=0,1,\cdots,N-1 \tag{6}$$

and

$$k_m=\begin{cases}1 & if\quad m\neq 0\\ 1/\sqrt{2}, & if\quad m=0\end{cases} \tag{7}$$

Because of the symmetrical properties of the cosine function, $[C]$ can be decomposed into sparse matrices. We show, in what follows, that if N is an even number, $[C]$ can be decomposed as:

$$\begin{bmatrix} C_N^{0,0} & C_N^{0,1} & \cdots & C_N^{0,N-2} & C_N^{0,N-1}\\ C_N^{1,0} & C_N^{1,1} & & C_N^{1,N-2} & C_N^{1,N-1}\\ \vdots & & & & \vdots\\ C_N^{N-2,0} & C_N^{N-2,0} & & C_N^{N-2,N-2} & C_N^{N-2,N-1}\\ C_N^{N-1,1} & C_N^{N-1,1} & \cdots & C_N^{N-1,N-2} & C_N^{N-1,N-1}\end{bmatrix}$$

$$=\begin{bmatrix} C_N^{0,0} & C_N^{0,1} & \cdots & C_N^{0,N/2-1} & 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0 & C_N^{1,N/2} & C_N^{1,N/2+1} & \cdots & C_N^{1,N-1}\\ \vdots & & & & \vdots & & & \vdots\\ C_N^{N-2,0} & C_N^{N-2,1} & \cdots & C_N^{N-2,N/2-1} & 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0 & C_N^{N-1,N/2} & C_N^{N-1,N/2+1} & \cdots & C_N^{N-1,N-1}\end{bmatrix}\begin{bmatrix} 1 & & & & & & & 1\\ & \ddots & & & & & \iddots & \\ & & 1 & 1 & & & & \\ & & -1 & 1 & & & & \\ & \iddots & & & & & \ddots & \\ -1 & & & & & & & 1\end{bmatrix}$$

$$=\frac{1}{\sqrt{2}}[D]\begin{bmatrix} I_{N/2} & \tilde{I}_{N/2}\\ -\tilde{I}_{N/2} & I_{N/2}\end{bmatrix}=\frac{1}{\sqrt{2}}[D][B] \tag{8}$$

where $[I_{N/2}]$ is an $N/2$ by $N/2$ identity matrix, $[\tilde{I}_{N/2}]$ is an $N/2$ by $N/2$ opposite identity matrix, $[D]$ is an N by N conversion matrix and $[B]$ is an N by N butterfly matrix.

If $[D]$ is used as a kernel transform matrix, a new discrete transform is formed. Since $[D]$ is block matrix, we call this new transform the block discrete cosine transform (BDCT).

4. PROPERTIES AND PERFORMANCE OF THE BDCT

A. The unitarity property

Let d_m denote the *i*th column vector in the matrix $[D]$, then it can be shown that the inner product of two such vectors is

$$\langle d_m,d_k\rangle=\begin{cases}2\sum_{n=0}^{N/2-1}C_N^{m,n}C_N^{k,n} & \text{if both } m \text{ and } n \text{ are even number}\\ 2\sum_{n=N/2}^{N}C_N^{m,n}C_N^{k,m}, & \text{if both } m \text{ and } n \text{ are odd number}\\ 0, & \text{Otherwise}\end{cases} \tag{9}$$

where

$$2\sum_{n=0}^{N/2-1}C_N^{m,n}C_N^{k,n}=\sum_{n=0}^{N/2-1}C_{N/2}^{m/2,n}C_{N/2}^{k/2,n}$$
$$=\sum_{n=0}^{P-1}C_P^{m/2,n}C_P^{k/2,n}=\left\langle g_{m/2}^P,g_{k/2}^P\right\rangle=\delta_{m/2,k/2}=\delta_{m,k} \tag{10}$$

and

$$2\sum_{n=N/2}^{N}C_N^{m,n}C_N^{k,m}=\sum_{n=N/2}^{N}C_N^{m,n}C_N^{k,m}+\sum_{n=N/2}^{N}C_N^{m,n}C_N^{k,m}$$
$$=\sum_{n=N/2}^{N}C_N^{m,N-n}C_N^{k,N-n}+\sum_{n=N/2}^{N}C_N^{m,n}C_N^{k,m}$$
$$=\sum_{n=0}^{N/2-1}C_N^{m,n}C_N^{k,n}+\sum_{n=N/2}^{N}C_N^{m,n}C_N^{k,n}=\left\langle g_m^N,g_k^N\right\rangle=\delta_{mk} \tag{11}$$

where g_m^N denotes the *m*th column vector of the N by N matrix $[C]$, $P=N/2$, and

$$\delta_{mk}=\begin{cases}1, \text{for } k=m\\ 0, \text{for } k\neq m\end{cases} \tag{12}$$

Combining (5), (6) and (7), we can now represent the unitarity property of the BDCT by

$$\langle d_m,d_k\rangle=\delta_{mk} \tag{13}$$

B. Energy packing efficiency

The energy packing efficiency (EPE) is used as a criterion that measures the effectiveness of a transform in data compression. It is defined as the energy proportion contained in the first *M* of *N* transform coefficients. Rao and Yip [16] showed that the EPE is equivalent to the ratio of the sum of the first *M* diagonal elements to the sum of all the *N* diagonal elements of the auto-covariance matrix in the transform domain. Let $[A]$ be the data covariance matrix and $[T]$ be the transform. Then, the covariance matrix in the transform domain $[A^T]$ is given by

$$[A^T]=[T][A][T]^{-1} \tag{14}$$

Thus, the EPE for the transform, by definition, is

$$EPE(M)=\sum_{j=1}^{M}A_{jj}^T\Big/\sum_{j=1}^{N}A_{jj}^T \tag{15}$$

One standard comparison the EPE for different transforms is based on the assumption that the random sequence is governed by a first-order Markov process with the adjacent correlation coefficient ρ specified. In Figure 1, comparisons are shown for the EPEs of the BDCT and DCT based on a Markov-1 signal of $\rho=0.9$ and *N=24*.

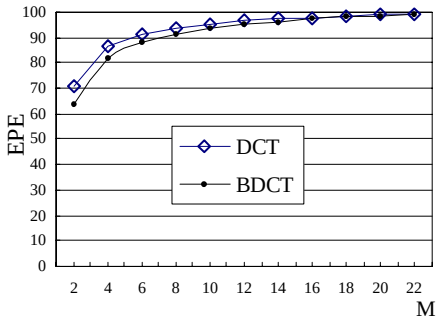

Figure 1. EPEs for DCT and BDCT

An inspection of Figure 1 reveals that the EPE of the BDCT is extremely close to that of the DCT when $M \geq 10$, which indicates the effectiveness of the BDCT in data compression.

In cepstral analysis used for speech recognition, the first cepstral coefficient (C_0) is often neglected, and the succeeding M coefficients are used as features. For this reason, we show in Figure 2, the EPE for both the BDCT and the DCT. One can see the efficiency of the BDCT in keeping information in cepstral analysis.

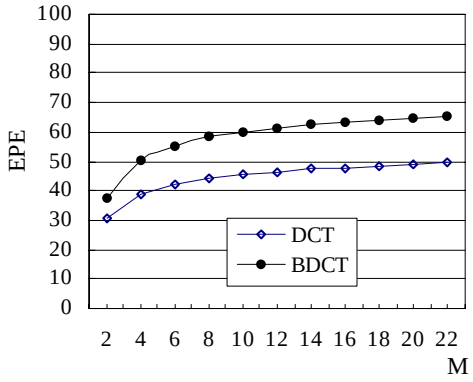

Figure 2. EPEs after removing the first coefficient

5. MEL FREQUENCY CEPSTRA USING THE BDCT

Mel frequency cepstrum coefficients (MFCCs) are perhaps the most widely used features for speech recognition. Conventionally, the MFCCs are computed by applying a DCT to the log mel filter bank energies. In this paper, we proposed to use the BDCT to replace the DCT in the estimation of the MFCCs. The resulting new MFCCs are called Block discrete cosine transform based MFCCs (BMFCCs). The BMFCCs haven been shown to have the following properties:

1. They may contain more information than the traditional MFCCs as the BDCT is shown to be able to preserve more information than DCT in cepstral analysis.
2. A sub-band processing strategy is implicit in the BMFCCs. To see this more clearly, the $M \times N$ conversion matrix which transforms the FEBs to cepstral coefficients can be written as follows without losing any information. (We assume *M* and *N* are even numbers).

$$[D]=\begin{bmatrix} C_N^{2,0} & C_N^{2,1} & \cdots & C_N^{2,N/2-1} & & & \\ C_N^{4,0} & C_N^{4,1} & \cdots & C_N^{4,N/2-1} & & & \\ \vdots & & & \vdots & & & \\ C_N^{M,0} & C_N^{M,1} & \cdots & C_N^{M,N/2-1} & & & \\ & & & & C_N^{1,N/2} & C_N^{1,N/2+1} & C_N^{1,N-1} \\ & & & & C_N^{3,N/2} & C_N^{3,N/2+1} & C_N^{3,N-1} \\ & & & & \vdots & & \vdots \\ & & & & C_N^{M-1,N-1} & C_N^{M-1,N-1} & C_N^{M-1,N-1} \end{bmatrix} \quad (16)$$

We can see that [*D*] is a block diagonal matrix. When it is applied to the FEBs, [*D*] automatically divides them into two sub-bands.

6. SPEECH RECOGNITION EXPERIMENTS

In this paper, we investigate the use of different features for speaker independent isolated speech recognition. The speech database used for this task is the ISOLET spoken letter database from OGI [11]. Here, the vocabulary consists of 26 English letters (A-Z). From this database, we take 90 utterances for each word from 45 male and 45 female talkers for training and 30 utterances for each word from 15 male and 15 female talkers (different from training talkers) for testing. The original speech was sampled at 16 kHz. We down-sample the speech to 8 kHz using a lowpass filter with a cutoff frequency of 3.5 kHz. The speech signal is analyzed every 12.5 ms with a frame width of 25 ms (with Hamming window and preemphasis).

To test the robustness of different feature sets with respect to noise, we directly add some noise to the speech signal in the test set. The training speech is kept clean. The noise signals used are from NOISEX database [17]. We down-sample the noise signal from 16 kHz to 8 kHz.

The recognition system uses a multi-mixture continuous density HMM framework. We use a 6-state continuous density HMM recognizer with probability density function approximated by a mixture of 5 multivariate normal distributions with diagonal covariance matrices.

The feature sets investigated include:

- FB MFCC (full-band MFCC): 12 MFCCs + 12 Δ MFCCs. The Mel frequency filter bank consists of 24 triangular filters.
- MB MFCC (multi-band MFCC): 12 multi-band MFCCs + 12 Δ multi-band MFCCs. We divide the full-band spectrum into 2 sub-bands: (0-1257 Hz) (1104-4000 Hz). Each subband is passed though a filter bank which contains 12 mel-scale triangle filters. The resulting 12 FBEs are converted to 6 MFCCs. We then concatenate these two sub-band MFCC vectors into a 12 dimensional sub-band MFCC vector.
- BMFCC: 12 BMFCCs + 12 Δ BMFCCs. The BMFCCs are computed from applying the BDCT to 24 mel frequency FBEs.

The recognition results for this experiment under three different noise conditions are shown in Tables 1, 2 and 3.

Feature set	Clean speech	30dB	20dB	15dB	10dB
FB MFCC	88.0	87.7	84.9	67.1	31.3
MB MFCC	87.8	87.5	87.1	77.4	41.7
BMFCC	89.5	89.6	88.0	75.4	35.1

Table 1. Speech recognition accuracy (in %) in speech noise condition (Average speech spectrum).

Feature set	Clean speech	30dB	20dB	15dB	10dB
FB MFCC	88.0	87.8	85.7	78.7	68.0
MB MFCC	87.8	87.8	86.9	80.0	69.1
BMFCC	89.5	89.4	89.4	82.6	73.5

Table 2. Speech recognition accuracy (%) in machine gun noise condition (Calibre 0.50, repeated).

Feature set	Clean speech	30dB	20dB	15dB	10dB
FB MFCC	88.0	88.0	87.6	87.1	84.7
MB MFCC	87.8	87.8	87.7	87.3	84.6
BMFCC	89.5	89.5	89.5	89.4	87.8

Table 3. Speech recognition accuracy (%) in car noise condition (Car-Volvo-340 120 km/h, 4th gear).

From these experiment results, we can make following observations:

1. The BMFCCs yield the best performance in clean speech and high SNR conditions. This demonstrates the superiority of the BDCT to the DCT in cepstral analysis for speech recognition.
2. All sub-band based features are more robust than the full-band MFCCs to three types of noise investigated.
3. BMFCCs are more robust to the machine gun noise and car noise than the multi-band MFCCs. While its robustness to speech noise is slightly poorer than that of the multi-band MFCCs. Tthe reason for this is not clear. We will perform further experiments with various speech databases and with more kinds of real noise before we draw a conclusion.

7. CONCLUSION

In this paper, we proposed a block DCT and applied it to the mel frequency cepstral analysis in speech recognition. We showed that a sub-band processing idea is implicit in the new features (BMFCCs). Experiment results based on a continuous density isolated speech recognizer revealed that the new MFCCs yield better recognition accuracy than full-band MFCCs in noise as well as in clean speech conditions.

BMFCCs were also compared with the multi-band MFCCs in terms of their recognition performance. The results showed that the BMFCCs were able to yield better performance in clean speech and various noisy speech environments. Thus, BMFCCs are more robust than the multi-band MFCCs under various noise conditions.

In this paper, experiments are only performed for small vocabulary isolated speech recognition tasks. Work is in progress to test the BMFCCs and various sub-band based front-ends for large vocabulary continuous speech recognition in clean as well as noise conditions.

REFERENCE

[1] D. S. Pallett, *et al*, "1996 Preliminary Broadcast News Benchmark", Proceedings of the 1997 DARPA Speech Recognition Workshop, International Conference Center Chantilly, Virginia, February 2-5, 1997

[2] D. S. Pallett, *et al*, "1997 Broadcast News Benchmark Test Results: English and Non-English", Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne Conference Resort, Lansdowne, Virginia, February 8-11, 1998.

[3] D. S. Pallett, *et al* "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures", Proceedings of the DARPA Broadcast News Workshop, Hilton at Washington Dulles Airport Herndon, Virginia, February 28-March 3, 1999.

[4] P. McCourt, S. Vaseghi and N. Harte, "Multi-Resolution Cepstral Features for Phoneme Recognition Across Speech Sub-bands," ICASSP'98, Seattle, USA, PP.557-560.

[5] S. Okawa, E. Bocchieri and A. Potamianos, "Multi-Band Speech Recognition in Noisy Environments," ICASSP'98, Seattle, USA, PP.641-644.

[6] A. Chen, S. Vaseghi and P. McCourt, "Transformation of Full-Band to Sub-band HMMS for Speech Recognition in Noisy Car Environments," ARSU'99, Keystone, Colorado, December 1999, Vol. 1, PP. 31-34.

[7] B. Doherty, S. Vaseghi and P. McCourt, "Linear Transformations in Sub-band Groups for Speech Recognition," EUROSPEECH'99, Budapest, Hungary, September, 1999, PP. 1359-1366.

[8] H. Bourlard and S. Dupont, "A mew ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," ICSLP'96, Philadelphia, October 1996.

[9] H. Bourlard and S. Dupont, "Subband-based Speech Recognition," ICASSP'97, PP. 1251-1254.

[10] S. Okawa, T. Nakajima and K. Shiria, "A Recombination Strategy for Multi-band Speech Recognition Based on Mutual Information Criterion," EUROSPEECH'99, PP. 603-606.

[11] K. K. Paliwal, "Decorrelated and Liftered Filter-Bank Energies for Robust Speech Recogntion," EUROSPEECH'99, PP. 85-88.

[12] R. Chengalvarayan, "Hierarchical Subband Linear Prediction Cepstral (HSLPC) Features for HMM-Based Speech Recognition," ICASSP'99, PP. 409-412.

[13] K. Yoshida K. Takagi and K. Ozeki, "Speaker Identification Using Subband HMMS," EUROSPEECH'99, PP. 1019-1022.

[14] S. Rao and W. A. Pearlman, "Analysis of Linear Prediction, Coding, and Spectral Estimation from Subbands," IEEE Trans. on Information Theory, Vol. 42, No. 4, 1996, PP. 1160-1178.

[15] A. D. Poularikas, "The Transforms and Applications Handbook," IEEE Press, 1995.

[16] K. R. Rao and R. Yip, "Discrete Cosine Transform: Algorithm, Advantages, Application," Academic Press, Inc, 1990.

[17] A. Varga, *et al*, "The Moise-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," DRA Speech Research Unit, St. Andrew's Rd., Malvern, Worcestershire, WR14 3PS UK.