THE DISCRETE COSINE TRANSFORM AND ITS SIGNIFICANCE IN AUDIO TECHNOLOGY AND ROOM ACOUSTICS

FF Li Manchester Metropolitan University, UK

1 INTRODUCTION

Over the past few years, the Discrete Cosine Transform (DCT) has attracted much attention in the research community of multimedia data encoding, audio and video signal compression and digital watermarking. Taking the advantage of its almost optimised coding gain and fast algorithms, the DCT and its variations such as the Modified Discrete Cosine Transform (MDCT), lapped transforms are found effective in image, audio and video signal compression applications and therefore are adopted in the JPEG, MPEG and H.26x standards. Apart from MPEG, audio and video compression and digital watermarking, the significance of the DCT in other audio and acoustic applications is less explored. This paper discusses the significance and potential applications of the DCT in context of audio technology and room acoustics.

A brief review and summary of the algorithms and properties of the DCT show that when dealing with stochastic processes, the DCT results in the concentration of signal energy and yields decorrelated coefficients. In particular, for a first order Markov process, the DCT gives very similar results to the optimal Karhunen-Loève Transform (KLT), but at a much reduced computational cost. For most random signals, not necessarily restricted to first order Markov processes, a long sequence can be fairly accurately coded with just a few DCT coefficients. As a result the DCT offers a high coding gain for most audio and acoustic signals such as music, speech and room impulse responses.

The KLT is just another name for the Principal Component Analysis (PCA) in statistics. Consequently, the DCT can be viewed as a fast algorithm for principal component estimation and the DCT encoding is actuarially a quasi-principal component encoding. Given the importance of the PCA in statistical signal processing, pattern recognition and the usefulness of the PCA in room impulse response modelling, the significance of the DCT in audio and acoustic applications becomes apparent.

Following the review of the DCT algorithms and the related properties, this paper will then focus on various potential audio and acoustic applications. In additional to signal compression and encoding, the DCT is found useful in modelling room impulse responses and implementing fast algorithms for room effects; Echo detection and cancellation can be made easier in the DCT domain; Wave Field Synthesis algorithms may be improved using the DCT; And last, but not least, the DCT can be used to approximate PCA and help extract acoustic features and parameters from received speech signals.

2 DISCRETE COSINE TRANSFORM AND ITS PROPERTIES

Derived from the famous Discrete Fourier Transform (DFT), the discrete cosine transform was first introduced in a short but important paper published in 1974[1]. Rao and Yip's monograph is probably the most complete and extensive treatment of the definition, mathematical background and potential applications of this simple yet attractive transform to this date[2].

With the prevalence of MPEG encoded audio and video programmes, DCT codecs are commonplace in home entertainment systems and Internet audio and videos. Most people have experienced DCT transformed sounds and images, but the DCT is still less well known than the Fourier transform to most researchers in the fields other than mathematics. A brief review of its definition, properties and the relations with other transforms would give insights into the potential of this transform.

2.1 Definitions

The discrete cosine transform on a real number series is a linear, orthogonal, invertible and purely real transform that maps the original number series onto its frequency domain. There are 12 slightly different DCTs in the literature but the commonly used is the one known as DCT–II, simply refereed to as DCT by most authors.

The DCT $X_c(k)$ of a real series *x(n) of length N is defined by*

$$X_c(k) = c(k)\sum_{n=1}^{N} x(n)\cos\frac{\pi(2n-1)(k-1)}{2N}; \qquad k \in [1, N] \tag{1}$$

and its inverse IDCT can be written as

$$x(n) = \sum_{k=1}^{N} c(k)X_c(k)\cos\frac{\pi(2n-1)k(k-1)}{2N}; \qquad n \in [1, N] \tag{2}$$

where

$$c(k) = \begin{cases} \sqrt{\frac{1}{N}}; & k = 1 \\ \sqrt{\frac{2}{N}}; & 2 \le k \le N \end{cases} \tag{3}$$

By the definition, the DCT has a cosine basis function; is orthogonal and yields a purely real frequency domain representation of the original time series. The above definitions read rather similar to the real part of DFT, but the difference between these two transforms stems from different assumptions: the DFT assumes that the time series is periodically continued with a period of *N*, whereas the DCT assumes that the series continued with its mirror image and then periodically continued with a period of *2N*. That is to say the DCT can be calculated from the real part of DFT on a double-length time series obtained by mirroring the original one. This gives one possible FFT based fast algorithm for the DCT. Such a fast DCT is typically achieved using pre- and post processors and an FFT routine [2].

2.2 Summary of properties and useful features

The DCT has four important properties:

1. For any DCT performed on a real number series, there always exists an inverse DCT.

2. The transform basis vector is orthogonal.

3. The Parseval's and Plancherel's Relations apply.

4. DCT spectrum ≠ Fourier spectrum! The DCT spectrum does not have a clear physical meaning. It contains Fourier spectrum and aliasing.

The DCT is known to work better than the DFT when dealing signals with strong correlation. A long sequence can be fairly accurately recovered from a few DCT coefficients. This is a useful property for applications requiring data reduction and therefore is used in JPEG, MPEG, and H.26x standards for image, video and audio signal compression.

2.3 Relations with PCA

In statistics and statistical signal processing, Principal Components Analysis (PCA) is a coordinate rotation approach to data reduction, intending to retain the significant characteristics of the original dataset or signals. The PCA adaptively chooses a new coordinate system according to the dataset such that the greatest variance by any projection of the dataset lies on the first axis, the second greatest variance on the second axis, and so forth. The first few principal components are retained but the latter principal components are discarded to achieve data reduction. The PCA has been commonplace in pattern recognition, because it reduces data dimensionality and retains characteristics of the dataset that contribute most to its variance. As a result, these characteristics are statistically important. However, it is worth noting that the actual effectiveness of PCA in feature extraction and pattern recognition depends on the applications. In signal processing, the PCA is also known as the Karhunen-Loève transform (KLT) in recognition of their work published about 70 years ago[3,4].

In essence, the PCA or KLT is an optimal linear transform for maximum variance. In stead of using a predefined basis function, the PCA or KLT uses a dataset dependent basis function obtained by eigenvectors of the covariance matrix of the dataset. Eigenspace decomposition is known to be a computationally intensive task. An alternative approach to eigenspace is the use of a kind of self-organising artificial neural network known as PCA network. The latter might have reduced computational complexity if only the first few principal components are needed. Nonetheless, both algorithms are time consuming and are not intended for real time processing of audio signals.

The attractiveness of the DCT is that although it uses a fixed basis function, it is asymptotically equivalent to the KLT for many real life random signals, including audio image and video[2]. In particular, it is an accurate approximation of first-order Gauss-Markov processes. Detailed proofs can be found in a recently published paper[5]. In practical uses, large blocks of data are normally needed for good approximation results.

2.4 Fast algorithms and hardware implementation

There exist dozens of fast algorithms for DCT. Typically, Fast DCTs (FDCTs) deploy other existing fast algorithms. For example FFT based FDCTs[2,6], Walsh-Hadamard transform based approaches[7], and the discrete Hartley transform (DHT) based approach[8] can all offer near real-time performance with a modern desktop computer. Most recently a new algorithm is developed to perform DCT without multiplication[9].

In addition to these fast algorithms, there are also several low cost hardware chips that implement DCT in real time, for example, the 1-D DCT chip developed by XiLinx Inc. and the LF3320 DCT chip from the Logic Devices Inc..

3 AUDIO ENCODING AND COMPRESSION APPLICATIONS

3.1 Audio compression

DCT based audio compression is adopted by the MPEG audio standard. The MP3 is probably the most prevalent audio application. The DCT is completely invertible and so does not perform lossy compression by itself, but it offers signal energy condensation. The lossy compression is achieved via quantisation and the use of perceptual models in MPEG compression. Audio signals are transformed into DCT domain. The DCT results in signal energy concentration. After quantisation in conjunction with psychoacoustic model, a large number of the DCT coefficients become zeros, which do not need to be coded or transmitted. MP3 codecs are not new to most audio and acoustic professionals. More details about DCT audio compression can be found in the literature and standards[10].

3.2 Digital watermarking

Digital watermarking techniques are used for copyright protection and ownership authentication purposes. Although the author would argue that DCT domain watermarking techniques are not necessarily an optimal solution to audio signals, they seem to be quite popular among many other alternatives[12, 13] and can be effective[14].

3.3 Auxiliary audio channels

Auxiliary audio channels may find applications in communications, audio signal transmission and sound reproduction. They can be used to enhance services or provide extra functionality. For example, Dolby 5.1 multi-channel scheme can significantly improve spatial impression and dialogue intelligibility. Multi-channel schemes may be achieved using multi-channel transmission/storage or special codecs. The former means more physical channels and the latter implies increased demand of channel capacity. Transmission of an extra channel normally requires some fundamental changes to existing equipment and systems from program production, dubbing, recording, and transmission to final reproduction. Creating an extra virtual channel in a standard compatible audio format has significant advantages and provides a cost-effective solution to additional functionality without resort to modification of existing systems. A pilot study has been carried out to embed an auxiliary speech channel in an ordinary sound track in DCT domain using psychoacoustic model and information-hiding techniques.

Audio signals of a sound track are compressed with the freed bandwidth being used to form an auxiliary channel. Narrow-band speech signals are embedded and hidden in the auxiliary channel in the DCT domain such that the distortion is maintained under the perception threshold. The so formed composite signals are transformed back to time domain and can be used as normal audio signals for listening, processing, recording and transmission. If needed, the embedded speech in the auxiliary channel can be decoded. Simulation and subjective tests have shown the effectiveness of the method. The developed method can be used for speech enhancement.

4 POTENTIAL APPLICATIONS TO ROOM ACOUSTICS

As summarised, there are three significant advantages of the DCT: decorelated coefficients, energy concentration in few real number coefficients and optimal KLT approximation. This makes the DCT useful in room acoustic problems, where signal compression, data reduction or feature extraction is required.

4.1 Wave Field Synthesis application

Wave Field Synthesis (WFS) is a sound field reproduction technique that uses loudspeaker arrays to accurately render sound field in a listening space. The WFS is intended to overcome the limitations of conventional multi-channel sound reproduction.

The WFS uses tens to several hundreds channels to drive a large number of loudspeaker arrays placed in a space to synthesize a desired sound field in a predefined listening area. The WFS is relatively straightforward to achieve good results in free fields. But in realistic listening spaces, the acoustic conditions of the spaces complicate the scenario. Reflections in such spaces distort the synthesized wave fields. Typical solution to this problem is to use feedback control of the wave field such that the actual wave field is continuously monitored and distortions caused by reflections minimised. In order to adaptively obtain the compensation filter banks for all channels, a large number of microphones are used and a matrix of compensation filter banks obtained[15]. This approach has several limitations: First, MIMO (multi-input multi-output) system is known to be difficult to control. The large number of loudspeakers and microphones further complicates the case. A huge filter bank matrix needs to be adaptively calculated. This can hardly achieve in real time and the time delay causes secondary complications. (2) Since the signals are all correlated, it is difficult to effectively calculate the required compensation filter bank using existing techniques [16]. Spors et al. has recently proposed a new method, which allows the complete profile of the wavefiled to be determined using microphones placed on the boundary of an enclosure. This new method is called wavefield transform[17]. It addressed the above two problems to some extends.

As an alternative, DCT might play a useful role in simplifying the algorithm and improve the performance of the WFS. The key issue here is to decouple the room transfer function matrix. Mathematically, this requires Singular Value Decomposition (SVD)[17]. As reviewed in the previous section, the KLT is based on the SVD. In other words, if optimal KLT is performed, SVD is done. The DCT is a good approximator for the KLT and is computationally efficient. So, if the DCT is applied, the transfer function matrix can be approximately de-coupled. The DCT seems to be a solution to the WFS problem.

4.2 Echo cancellation and delay detection

Echo detection and cancellation has been a long-standing problem not only in room acoustics but also in telecommunications, seismic, ultrasonic signal processing and many other areas. LMS adaptive filtering techniques have found widespread applications in acoustic and transmission echo cancellation in telecommunications. However, due to the complexity of speech signals, the LMS algorithm is not sufficiently effective[18].

Abouchakra and Kabal proposed a more effective echo cancellation and delay detection algorithm – the DCT-LMS algorithm[18]. The algorithm has two stages: a DCT transform pre-processor and a LMS filter. A moving window is applied to input speech signals. The windowed portion of the signal is passed through a tapped delay line, and then transformed into DCT domain. The second stage is a typical LMS adaptive filter. The adaptive filter will track the DCT of the room impulse response by seeking to match exactly the DCT coefficients of the impulse response. Their results show that in the DCT domain the adaptive filter outperforms the traditional LMS echo cancellation filters and can more effectively detect the delay time of the major echoes and cancel them. Although the filter taps in their implementation is not short enough to resolve details of room impulse responses, the accurate detection of major reflections, i.e. the discrete peaks in impulse responses is a significant step towards blind estimation of room impulse responses from arbitrary speech excitation.

4.3 Modelling room impulse responses

The DCT can be used to code and compress impulse responses. Since an impulse response of room contains a lot of high frequency random contents, it can be significantly compressed without loosing its feature. Figure 1 shows a 2:1 compression of a simulated impulse response. Visual inspection can hardly find any difference between the original and compressed versions. Major parameters and backward integration results from these two versions are found identical.

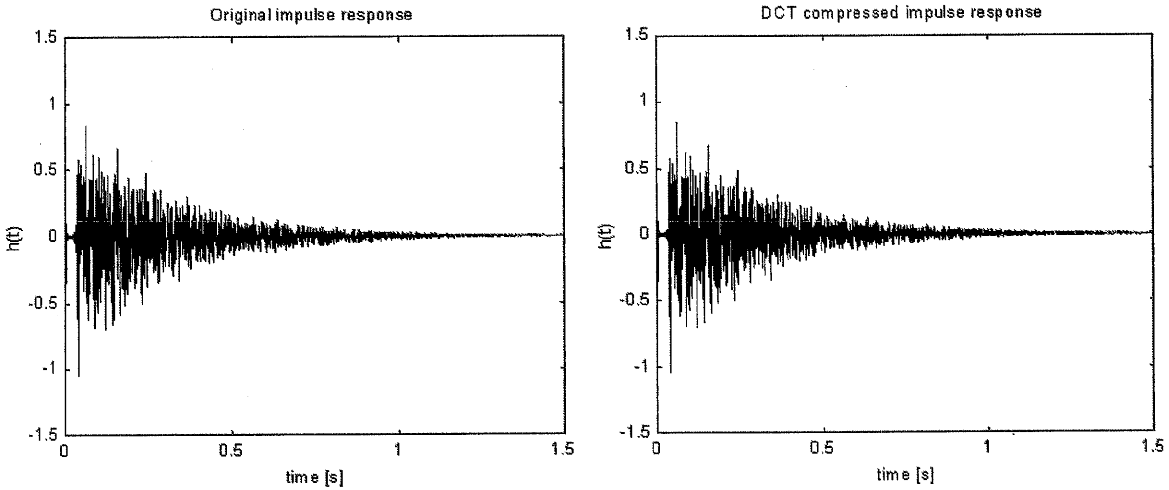

Figure 1. Original and DCT compressed impulse responses

Typically, 2:1 compression can retain all details of original impulse responses. A 7:1 compression ratio is found to give almost identical Schroeder backward integration plots. For reverberation time information only, the compression ratio can be further increase to above 10:1. As a result, the DCT can be used to model room impulse responses.

DCT impulse response encoding and compression can also be applied to multi-channel auralization or virtual reality sound effects, where a large number of impulse responses need to be processed or variable impulse responses are considered. Typical examples include auralization for virtual spaces and wave field synthesis.

4.4 Acoustic parameters estimation

Estimation of room acoustic parameters from naturalistic signals such as music and speech has significant applications in room acoustic measurement. It facilitates occupied measurement to obtain realistic acoustic profiles of a space. Parameters such as Reverberation Time (RT) and Speech Transmission Index (STI) can be accurately estimated from received speech signals using machine learning algorithms[19,20]. They can also be estimated to a reasonable accuracy without the knowledge of the speech sources using a PCA pre-processor and a machine learning routine such as an artificial neural network[21]. Such a hybrid model for blind estimation of STI is illustrated in Figure 2. In this model the PCA of the Hilbert envelope of speech signals functions as an automated feature selection stage and is implemented using a self-organising PCA network. The drawback of this approach is that the PCA may not converge to principal component subspace in a single trial. Several restarts are often needed. This makes the estimation process quite time consuming. For room acoustic parameter measurement, fast algorithm is not essential but a bonus, since it uses naturally occurring sounds anyway. But for some other applications, for example monitoring the speech intelligibility of a telecommunications channel and then equalizing the channel, real-time performance is essential.

Since the DCT approaches the KLT or PCA for long series, the DCT is used to replace the PCA network. Some simulations were carried out. The results are not surprising. The DCT performed almost as well as the PCA in this case as the feature selector. The extra benefit is much reduced computing time. The PCA-ANN method shows a maximum prediction error of 0.087 in STI, the DCT-ANN method has a maximum prediction error of 0.094. This experiment supports the aforementioned characteristics of the DCT – DCT gives results similar to the KLT (PCA) under certain conditions.

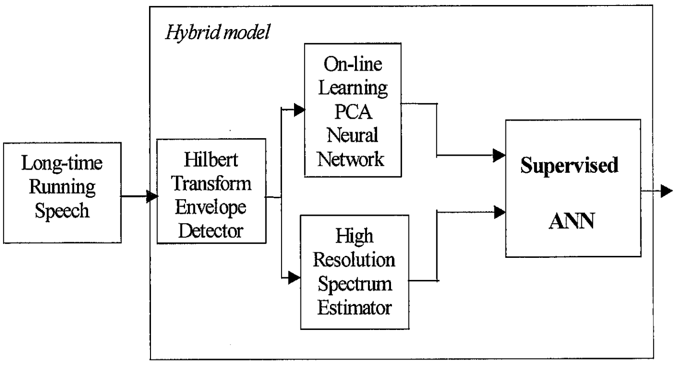
Hybrid model
Long-time Running Speech
Hilbert Transform Envelope Detector
On-line Learning PCA Neural Network
High Resolution Spectrum Estimator
Supervised ANN

Figure 2. PCA-ANN approach to blind STI estimation

5 CONCLUDING REMARKS

This paper has reviewed some important properties and useful features of the discrete cosine transform in the context of audio technology and room acoustics. The DCT approximates the KLT, (also known as PCA) for first order Markov processes and shows good computational efficiency when compared with the KLT. The PCA is an important tool for signal encoding, data reduction and feature selection. Consequently, the DCT, as a computational effective approximation of the KLT, may find many applications in audio and room acoustics. In particular, the DCT can potentially mitigate the technical difficulties in audio information hiding, wave field synthesis, impulse response encoding, fast auralization, echo detection and cancellation and blind room acoustic parameter estimation. Given the relevance and potential applications, the DCT may play a significant role not only in audio and video signal compression but also in many other audio and room acoustic areas.

6 REFERENCES

1. N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Trans. Comput.*, vol. C-23, pp. 90–93, Jan. (1974).
2. K. R. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, and Applications. New York: Academic. (1990).
3. M. M. Loève, “Fonctions Aleatories de Seconde Ordre”, in Process Stochastiques et Movement Brownien (P. Levt, ed.), Hermann, Paris. (1948).
4. K. Karhunen, “Ueber lineare Methoden”, in der Wahrscheinlichtskeitsrechnung. Annals Academy Science Fennicae Series A. I. Vol. 37. (1947).
5. J. M. F. Moura and M. G. S. Bruno, “DCT/DST and Gauss-Markov fields: conditions and equivalence”, *IEEE trans. on signal processing*, Vol. 46, No. 9, pp. 2571-2574. (1998).
6. J. Makhoul, “A fast cosine transform in one and two dimentions,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 27–34. (Feb. 1980).
7. S. Venkataraman, V. Kanchan, K. R. Rao and M. Mohanty, “Discrete transform via the Walsh-Hadamard transform,” *Signal Processing*, vol. 14, no. 4, pp. 371–382. (June 1988).

8. H. Malvar, "Fast computation of the discrete cosine transform and the discrete Hartley transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1484–1485. (Oct. 1987).
9. T. D. Tran, "The BinDCT: Fast Multiplierless Approximation of the DCT", *IEEE Signal precessing Letters*, Vol. 7, No. 6, pp. 141- 144. (2000).
10. P. Noll, "MPEG digital audio coding standards," in The Digital Signal Processing Handbook, V. Madisetti and D. B. Williams, Eds. Boca Raton, FL: CRC. (1998).
11. I. Yeo,and H. J. Kim, "Modified Patchwork Algorithm: A Novel Audio Watermarking Scheme", in Proceedings of International Conference on Information Technology: Coding and Computing (ITCC '01), Las Vegas. USA. (April 2001).
12. I. J. Cox and M. L. Miller, "The first 50 years of the electronic watermarking", *Journal of Applied Signal Processing*, Vol. 2, pp. 126-132. (2002).
13. M.D. Swanson, M. Kobayashi and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies", Proceedings of the IEEE, Vol. 86, No. 6, pp. 1064-1087. (1998).
14. J. T. Park and K. H. Rhee, "An implementation on the digital audio watermarking for high quality audio", in CD proceedings of the 2002 International Conference on Circuits/Systems Computers and Communications, Thailand. (July 2002).
15. L.D. Fielder. "Practical limits for room equalization", 111th AES Convention, preprint No.5481, New York. USA. (Sept. 2001).
16. M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation: an overview of the fundamental problem", *IEEE Signal Processing Letters*, 2(8), pp.148-151. (August 1995).
17. S. Spors, H. Buchner, and R Rabenstein, "Efficient Active Listening Romm Compensation for Wave Field Synthesis ", Preprint No. 6119, 116th AES convention, Berlin, Germany. (May 2004).
18. R. Abouchakra and P. Kabal, "Delay Estimation for Trasform Domain Acoustical echo Cancellation", Proc. European Conf. Speech Commun., Technology Budapest, Hungary, pp. 2539-2542. (Sept. 1999).
19. T. J. Cox, F. Li, and P. Darlington, "Extraction of room reverberation time from speech using artificial neural networks," *Journal of AES*, Vol. 49, No. 4, pp. 219-230. (2001).
20. F. F. Li and T. J. Cox, "Speech transmission index from running speech: A neural network approach," *Journal of Acoust. Soc. Am.*, Vol. 113, Issue 4, pp.1999-2008. (2003).
21. F. F. Li and T. J. Cox, "A neural network for blind identification of speech transmission index", Proceedings of IEEE ICASSP2003', V. II, pp. 757-760. (2003).